

Automatic Analysis of Naturalistic Hand-Over-Face Gestures

MARWA MAHMOUD, TADAS BALTRUŠAITIS, and PETER ROBINSON,
University of Cambridge

One of the main factors that limit the accuracy of facial analysis systems is hand occlusion. As the face becomes occluded, facial features are lost, corrupted, or erroneously detected. Hand-over-face occlusions are considered not only very common but also very challenging to handle. However, there is empirical evidence that some of these hand-over-face gestures serve as cues for recognition of cognitive mental states. In this article, we present an analysis of automatic detection and classification of hand-over-face gestures. We detect hand-over-face occlusions and classify hand-over-face gesture descriptors in videos of natural expressions using multi-modal fusion of different state-of-the-art spatial and spatio-temporal features. We show experimentally that we can successfully detect face occlusions with an accuracy of 83%. We also demonstrate that we can classify gesture descriptors (*hand shape*, *hand action*, and *facial region occluded*) significantly better than a naïve baseline. Our detailed quantitative analysis sheds some light on the challenges of automatic classification of hand-over-face gestures in natural expressions.

Categories and Subject Descriptors: I.2.10 [Vision and Scene Understanding]: Video Analysis

General Terms: Affective Computing, Body Expressions

Additional Key Words and Phrases: Hand-over-face occlusions, face touches, hand gestures, facial landmarks, histograms of oriented gradient, space-time interest points

ACM Reference Format:

Marwa Mahmoud, Tadas Baltrušaitis, and Peter Robinson. 2016. Automatic analysis of naturalistic hand-over-face gestures. *ACM Trans. Interact. Intell. Syst.* 6, 2, Article 19 (July 2016), 18 pages.
DOI: <http://dx.doi.org/10.1145/2946796>

1. INTRODUCTION

Over the past few years, there has been an increasing interest in machine understanding and recognition of people's affective and cognitive mental states, especially based on facial expression analysis. One of the major factors that limits the accuracy of facial analysis systems is hand occlusion. People often hold their hands near their faces as a gesture in natural conversation. As many facial analysis systems are based on geometric or appearance based facial features, such features are lost, corrupted, or erroneously detected during occlusion. This results in an incorrect analysis of the person's facial expression. Although face touches are very common, they are under-researched, mostly because segmenting of the hand on the face is very challenging, as face and hand usually have similar colour and texture. Detection of hand-over-face

The research leading to these results received partial funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant No. 289021 (ASC-Inclusion). We also thank Yousef Jameel and Qualcomm for providing funding as well.

Authors' address: The Computer Laboratory, 15 JJ Thomson Avenue, Cambridge CB3 0FD, United Kingdom; emails: {Marwa.Mahmoud, Tadas.Baltrusaitis, Peter.Robinson}@cl.cam.ac.uk.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 2016 ACM 2160-6455/2016/07-ART19 \$15.00

DOI: <http://dx.doi.org/10.1145/2946796>

occlusion can significantly improve facial landmark detection and facial expression inference systems.

However, hand-over-face occlusions are not just noise that needs to be removed. There is a growing body of work on the importance of body movements and gestures as significant visual cues that complement facial expressions in affect analysis [de Gelder 2009; Aviezer et al. 2012]. Body gestures were successfully utilised in automatic detection of human mental states [Bernhardt and Robinson 2007; Castellano et al. 2007]. For a survey on affective body expression perception and recognition, the reader is referred to Kleinsmith and Bianchi-Berthouze [2013]. Spontaneous self-touches or self-grooming gestures are believed to be related to formulating thoughts, information processing, and emotion regulation [Freedman 1977; Barroso and Feld 1986; Grunwald et al. 2014]. Specifically, recent studies suggest that hand-over-face gestures can serve as an additional valuable channel for multi-modal affect inference for cognitive mental states [Mahmoud and Robinson 2011]. These studies emphasise the need not only for an occlusion detection system but also for a way to describe the gesture in terms of a set of quantitative descriptors that can be automatically detected.

Moreover, automatic detection of these gesture descriptors can provide tools for experimental psychologists who study gesture—especially face touches—to detect and quantify these gestures automatically, instead of the common practice of manual coding. To date, there is no available automatic detection system that serves these purposes.

In this article, we present an analysis of hand-over-face gestures in a naturalistic video corpus of complex mental states. We define three hand-over-face gesture descriptors, namely *hand shape*, *hand action*, and *facial region occluded*, and propose a methodology for automatic detection of face occlusions in videos of natural expressions.

We treat the problem as two separate tasks: detection of hand occlusion and classification of hand gesture descriptors. The main contributions of this article are as follows:

- (1) Proposing a multi-modal fusion approach to detect hand-over-face gestures in videos of natural expressions, based on state-of-the-art spatial and spatio-temporal appearance features.
- (2) Proposing the first approach to automatically code and classify hand-over-face gesture descriptors, namely *hand shape*, *hand action*, and *facial region occluded*.
- (3) Demonstrating that multi-modal fusion of spatial and spatio-temporal features outperforms single modalities in all of our classification tasks.

We start by discussing the related work in Section 2. We present the details of gesture coding and dataset used in Section 3. We then present our proposed approach (illustrated in Figure 1), starting by the feature extraction in Section 4 followed by the experimental evaluation in Section 5. Conclusions and future directions are presented in Section 6.

2. RELATED WORK

There have been several previous attempts to detect and deal with occlusion in face area. Two such examples come from work done by Yu et al. [2013] and Burgos-Artizzu et al. [2013]. In both pieces of work, the authors concentrated on building a facial landmark detector that is robust to various occlusions. They achieved this by explicitly recognising occluded landmarks of the face and using that information to detect the visible landmarks more robustly.

Yu et al. [2013] evaluated their approach on static images of faces. They did not report occlusion detection or hand segmentation results. They just reported face alignment error rates in the presence of different types of face occlusion. Burgos-Artizzu et al. [2013] reported occlusion detection precision/recall curve for specific facial landmarks. Selected threshold for occlusion classification was reported at 0.8 precision and

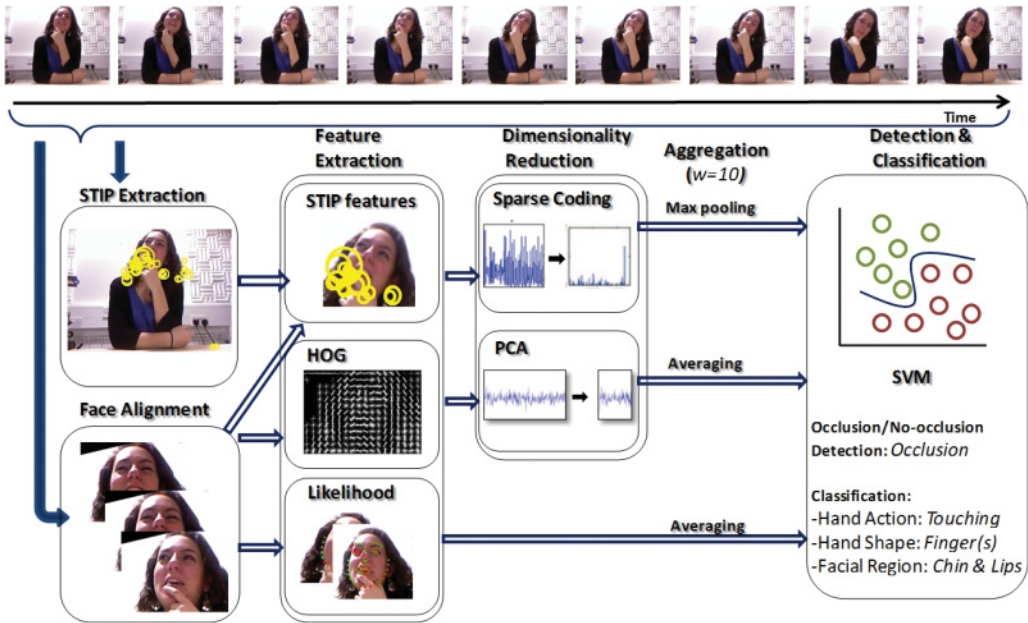


Fig. 1. Overview diagram shows the main steps in our approach.

0.4 recall values. Both pieces of work concentrated on facial occlusion in general and not specifically on hand-over-face occlusion. Furthermore, the authors were interested in detecting occluded facial landmarks, which are not necessarily semantically meaningful. Both of the approaches lead to better landmark alignment in the presence of occlusion.

Another notable example is the work done by Hotta [2004], which proposes a method for more robust face detection in presence of partial occlusion by using Support Vector Machines (SVM) with local kernels. Again, Hotta [2004] did not provide any quantitative results for the face occlusion detection. It is worth mentioning that none of these mentioned authors distinguish among types of occlusion and make no special analysis of hand-over-face occlusions.

There have been a few attempts [Smith et al. 2007; Mahmoud et al. 2009] to detect the hand when it occludes the face. Both studies evaluated their approaches on posed sets of face images that mostly included frontal faces with very limited illumination and head motion variations. Their approaches were based on the assumption that the face is the only object in the image as no face detection technique was employed.

A great deal of work has been published recently using depth captured from RGB-D sensors for hand shape detection in sign language [Keskin et al. 2011; Kurakin et al. 2012; Dong et al. 2015], gesture recognition for control gestures [Ren et al. 2011; Yang et al. 2012; Suarez and Murphy 2012], and general hand pose estimation [Poier et al. 2015]. Unlike hand-over-face gestures, in these scenarios one can assume that the hand is the closest body part to the depth sensor and is sufficiently far from the face to allow for segmentation.

Moreover, Gonzalez et al. [2012] use colour and edge information captured from RGB images to track and segment the hand during hand-over-face occlusion in sign language. They assume that the hand is mostly in front of the face, which is a fair assumption in the context of sign language detection. When the hand is in front of the face, some differences in the colour and edges of the hand are preserved as the



Fig. 2. Sample frames from videos in the dataset Cam3D showing examples of face touches present in the dataset [Mahmoud et al. 2011]. Note the challenging—close to natural—recording settings like inconsistent lighting conditions and strong head rotations.

hand edges are not fully merged with the face edges, especially with the existence of shadows. This assumption does not hold in the naturalistic hand-over-face occlusions that we are interested in.

Grafsgaard et al. [2012] use surface propagation from depth images to detect two hand-to-face gestures (one hand touching face and two hands touching face) in a computer-mediated tutoring environment. In contrast to previous work, our work presents a more detailed and comprehensive classification of hand-over-face gesture descriptors, which is not tackled in any of the previous studies.

3. CODING OF HAND-OVER-FACE GESTURES

Serving as a first step in automatic classification, we coded hand-over-face gestures using a set of descriptors. In this section, we describe the choice of the dataset, the coding schema, the labelling, annotation assessment, and how we generate the ground-truth labels that are used in our machine learning experiments.

3.1. Dataset

The first challenge was to find a corpus of videos of natural expressions. Most of the work on affect analysis focuses on the face, so most of the publicly available natural datasets also focus on faces with limited or no occlusion. Since we are also interested in the temporal aspect of the hand gesture, corpora of still photographs were not useful. The publicly available Cam3D corpus [Mahmoud et al. 2011] has natural expressions and does not restrict the video collection to faces. It includes upper body videos that have hand-over-face occlusions in around 25% of the videos. The expressions in Cam3D are elicited as part of an emotion elicitation experiment, which implies that the hand gestures expressed are most likely to be part of expression of emotion. We are interested in detecting such potentially informative gestures. Figure 2 shows examples of face touches present in the dataset Cam3D, showing the challenging—close to natural—recording settings like inconsistent lighting conditions and strong head rotations.

In Cam3D, segmentation is event based, so each video segment contains a single action. The dataset has 192 video segments that contain hand-over-face occlusions.

These videos come from nine participants with mean duration of each video being 6s. We used all of the occluded videos. For balance, we also randomly selected another 173 video segments from the Cam3D dataset that do not contain face occlusions. The no-occlusion videos were selected to contain the same nine participants while keeping the number of samples for each participant as balanced as possible. This led to a set of 365 videos in total.

3.2. Labelling

In order to proceed to automatic detection, we needed to code the hand-over-face occlusions present in the dataset. This requires a taxonomy of hand gestures that can form the basis of a set of descriptors.

There have been a few attempts to code hand gestures. Ip et al. [1998] developed a Hand Action Coding System (HACS) that is based on the anatomy of the hand to be used for hand synthesis. This system was too detailed to be used for our coding as it was anatomical rather than descriptive of the hand gesture.

Inspired by the coding schema provided by Mahmoud and Robinson [2011], we coded the gestures in terms of *hand shape*, *hand action*, and *facial region occluded*.

Labelling was carried out using the Elan video annotation tool [Lausberg and Sloetjes 2009]. Two expert coders (researchers in our research group) were instructed to label the videos given the following instructions:

—Hand Action: coded as one label for the whole video according to the action observed in the majority of the frames.

Labels are as follows: (1) **Touching**—if the hand is static while touching the face.

(2) **Stroking/tapping**—repetitive motion of the hand on the face. (3) **Sliding**—any other hand motion that is not repetitive.

—Hand Shape: coded as one label per frame. It describes the shape of the hand on the face in a specific frame. Labels are mutually exclusive, that is, one label is permitted per frame.

Labels are as follows: (1) **Fingers** or any separate fingers. (2) **OpenHand(s)** or palm(s). (3) **ClosedHand(s)** or a fist shape. (4) **HandsTogether**—tangled hands.

—Facial Region Occluded: coded as one (or multiple) labels per frame (labels are not mutually exclusive). It describes the face area covered—or partially covered—by the hand during occlusion.

Labels are as follows: (1) **Forehead**. (2) **Eye(s)**. (3) **Nose**. (4) **Cheek(s)**. (5) **Lips**. (6) **Chin**. (7) **Hair/ear**.

Figure 3 shows sample frames corresponding to different coding categories.

3.3. Coding Assessment and Refinement

To assess the coding schema and gain confidence in the labels obtained, we calculated inter-rater agreement between the two expert annotators using time-slice Krippendorff's alpha [Krippendorff 2004], which is widely used for this type of coding assessment because of its independence from the number of assessors and its robustness against imperfect data [Hayes and Krippendorff 2007]. We got a Krippendorff's alpha coefficient of 0.92 for *hand action*, 0.67 for *hand shape*, and an average alpha coefficient of 0.56 for *facial region occluded* (forehead 0.69, eye(s) 0.27, Nose 0.45, cheeks 0.65, lips 0.73, chin 0.83, hair/ear 0.25). All the classes had moderate agreement or above except for the following facial regions: eyes, nose, and hair/ear. When we explored the reason of the disagreement in these categories, this was mostly because very few samples were available of these categories in the dataset, for example: eyes, forehead, and hair/ear regions had only 25, 100, and 10 frame samples respectively, that is, less than 0.2% of the total number of frames in total. We decided to exclude



Fig. 3. Sample frames for different categories in the hand-over-face coding scheme for three gesture descriptors: hand action, hand shape, and facial region occluded.

these categories (mostly upper face area) in the machine-learning step, as it was unfair to try to learn and classify these categories automatically when the human annotators failed to agree.

Due to the nature of our unbalanced dataset, some labels had very few samples. In the classification stage, we decided to aggregate some of the groups together. The nose region was combined with the cheek region as one descriptor of the middle face region. For the *hand action* descriptor, we combined sliding, stroking and tapping into a single group representing non-static hand gesture, that is, any type of motion. Figure 4 illustrates the final refined coding scheme that was used in our subsequent analysis.

4. FEATURE EXTRACTION

The first building block of our approach is feature extraction. We chose features that can effectively represent hand gesture descriptors that we want to detect. Therefore, we extract spatial features, namely Histograms of Oriented Gradients (HOG) [Dalal and Triggs 2005] and facial landmark alignment likelihood [Baltrušaitis et al. 2013]. Moreover, having the detection of hand action in mind, we also extract Space Time Interest Points (STIP) [Laptev 2005] that combine spatial and temporal information. For HOG and STIP features, dimensionality reduction of features is then applied to obtain a more compact feature representation.

4.1. Space-Time Features

Local space-time features [Laptev 2005; Laptev et al. 2008; Dollár et al. 2005] have become popular motion descriptors for action recognition [Poppe 2010]. They provide compact and abstract representations of patterns in an image. Recently, they have been used by Song et al. [2013] to encode facial and body microexpressions for emotion detection. They were particularly successful in learning the emotion valence dimension as they are sensitive to global motion in the video [Song et al. 2013]. Our methodology


Hand Action	Static			Dynamic				
Hand Shape	Fingers		Open hand		Closed hand		Two hands	
Facial Region Occluded	Chin			Lips			Middle face (Cheeks & Nose)	

Fig. 4. The refined coding scheme for hand-over-face gesture descriptors.

for space-time interest points feature extraction and representation is based on the approach proposed by Song et al. [2013].

STIP capture salient visual patterns in a space-time image volume by extending the local spatial image descriptor to the space-time domain. Obtaining local space-time features is a two step process: STIP detection followed by feature extraction. Wang et al. [2009] reports that using the Harris3D interest point detector followed by a combination of the HOG and the Histogram of Optical Flow (HOF) feature descriptors provide good performance. Thus, we use the Harris3D detector with HOG/HOF feature descriptors to extract local space-time features. As we are interested in the face area, we use the face alignment input to crop the STIP features and discard any extracted points outside the face region.

The STIP box in the overview diagram in Figure 1 shows how the hand motion is captured by the space-time features (denoted by the yellow circles in the diagram).

The local space-time features extracted are dense as they capture micro-expressions. Since we are interested in more semantic feature representation, we use sparse coding to represent them so only few salient features are recovered, that is, features that appear most frequently in the data. Thus, we learn a codebook of features and use it to encode the extracted features in a sparse manner.

The goal of sparse coding is to obtain a compact representation of an input signal using an over-complete codebook, that is, the number of codebook entries is larger than the dimension of input signal so only a small number of codebook entries are used to represent the input signal. Given an input signal $\mathbf{x} \in \mathbb{R}^N$ and over-complete codebook $\mathbf{D} \in \mathbb{R}^{N \times K}$, $K \gg N$, we find a sparse signal $\alpha \in \mathbb{R}^K$ that minimises the reconstruction error,

$$\min_{\alpha \in \mathbb{R}^K} \frac{1}{2} \|\mathbf{x} - \mathbf{D}\alpha\|_2^2 + \lambda \|\alpha\|_1, \quad (1)$$

where the first term in this equation measures reconstruction error and the second term is the L_1 regularisation that encourages the sparsity of vector α . λ controls the

relative importance of the two terms so we have α containing few non-zero linear coefficients compared to the codebook \mathbf{D} , which leads to the best approximation of \mathbf{x} .

In our work, we learn the codebook \mathbf{D} from our data, that is, the extracted space-time features $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M\}$,

$$\min_{\mathbf{D}} \frac{1}{M} \sum_{i=1}^M \min_{\alpha_i} \frac{1}{2} \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1. \quad (2)$$

The optimisation problem is convex in \mathbf{D} with $\mathbf{A} = [\alpha_1, \alpha_2, \dots, \alpha_M]$ fixed and in \mathbf{A} with \mathbf{D} fixed but not in both at the same time [Mairal et al. 2010]. Thus, it can be solved using online learning [Mairal et al. 2010] by alternating the two convex optimisation problems. Once the codebook is learned, we can use it to encode each space-time feature \mathbf{x} into α by solving Equation (2).

From each frame we obtain different number of local space-time features (and corresponding sparse codes). These features need to be aggregated to obtain a vector of a fixed dimension to be suitable for our classification step. Averaging or max pooling are typical ways of doing this. In our work, we use max pooling as it provides a representation that is more resilient against image transformations and noise [Yang et al. 2009; Song et al. 2013]. The max-pooling operation is defined as follows:

$$\mathbf{z} = [\max_{i=1\dots M_v} |\alpha_{i,1}|, \max_{i=1\dots M_v} |\alpha_{i,2}|, \dots, \max_{i=1\dots M_v} |\alpha_{i,K}|], \quad (3)$$

where M_v is the number of sparse codes associated with a given space-time volume v .

To obtain a more compact representation of the features and to speed up processing time, we aggregate the space-time features (and their corresponding sparse codes) over a window $w = 10$ frames. This step is explained in Section 5.1.

4.2. Facial Landmark Detection: Likelihood

Facial landmark detection plays a large role in face analysis systems. In our case it is important to know where the face is in order to compute HOG appearance features around the facial region and to remove irrelevant STIP features.

We employ a Constrained Local Neural Field (CLNF) [Baltrušaitis et al. 2013] facial landmark detector and tracker to allow us to analyse the facial region for hand-over-face gestures. CLNF is an instance of a Constrained Local Model (CLM) [Cristinacce and Cootes 2006] that uses more advanced patch experts and optimisation function. We use the publicly available CLNF implementation [Baltrušaitis et al. 2013].

In summary, the model works by first detecting a face region of interest (ROI) in an image and then finding the most likely arrangement of facial landmarks in that ROI. This is done by evaluating the likelihood of each landmark individually using a *local* patch expert and by guiding the updated locations using a *global* shape model that describes possible arrangements of landmarks. This process is repeated for every frame in a video.

The CLM model we use can be described by parameters $\mathbf{p} = [s, \mathbf{R}, \mathbf{p}, \mathbf{t}]$ that can be varied to acquire various instances of the model: the scale factor s ; object rotation \mathbf{R} (first two rows of a three-dimensional (3D) rotation matrix); 2D translation \mathbf{t} ; a vector describing non-rigid variation of shape \mathbf{p} . The point distribution model (PDM) is as follows:

$$\mathbf{x}_i = s \cdot \mathbf{R}(\bar{\mathbf{x}}_i + \Phi_i \mathbf{p}) + \mathbf{t}. \quad (4)$$

Here $\mathbf{x}_i = (x, y)$ denotes the 2D location of the i th feature point in an image, $\bar{\mathbf{x}}_i = (X, Y, Z)$ is the mean value of the i th element of the PDM in the 3D reference frame, and the vector Φ_i is the i th eigenvector obtained from the training set that describes



Fig. 5. An example of patch expert responses in presence of occlusion. Green shows high likelihood values, while red means low likelihoods.

the linear variations of non-rigid shape of this feature point, and the vector Ψ_i is the i th eigenvector that describes the linear variations of non-rigid shape.

In CLM (and CLNF) we estimate the maximum *a posteriori* probability (MAP) of the face model parameters \mathbf{p} :

$$p(\mathbf{p}|\{l_i = 1\}_{i=1}^n, \mathcal{I}) \propto p(\mathbf{p}) \prod_{i=1}^n p(l_i = 1|\mathbf{x}_i, \mathcal{I}), \quad (5)$$

where $l_i \in \{1, -1\}$ is a discrete random variable indicating if the i th feature point is aligned or misaligned, $p(\mathbf{p})$ is the prior probability of the model parameters \mathbf{p} , and $\prod_{i=1}^n p(l_i = 1|\mathbf{x}_i, \mathcal{I})$ is the joint probability of the feature points \mathbf{x} being aligned at a particular point \mathbf{x}_i , given an intensity image \mathcal{I} (see Section 4.2).

We employ a common two-step CLM fitting strategy [Cristinacce and Cootes 2006; Saragih et al. 2011]; performing an exhaustive local search around the current estimate of feature points leading to a response map around every feature point and then iteratively updating the model parameters to maximise Equation (5) until a convergence metric is reached. For fitting we use Non Uniform Regularised Landmark Mean-Shift [Baltrušaitis et al. 2013].

As CLNF is a local optimisation approach, it relies on initial face detection. However, few face detectors are suitable for the task in the presence of occlusion. In our work, we used a Deformable Parts Model– (DPM) based Zhu and Ramanan [2012] face detector to initialise landmark detection and tracking. The subsequent frames were initialised using the previous frames estimate, only requiring to run the DPM detector multiple times in a video: to initialise and to reinitialise when tracking fails. It would have been possible to use the DPM to detect landmarks in every video frame; however, it is not as accurate as dedicated landmark detectors such as CLNF [Baltrušaitis et al. 2013] and is too slow to be used for video analysis.

In CLM patch experts are used to calculate $p(l_i = 1|\mathbf{x}_i, \mathcal{I})$, which is the probability of a feature being aligned at point \mathbf{x}_i (Equation (4)). As a probabilistic patch expert we use a Continuous Conditional Neural Field regressor [Baltrušaitis et al. 2013], which given a local $n \times n$ image patch centered around current landmark estimate predicts the alignment likelihood.

The likelihood response from the patch expert will be low when it is either not aligned or the landmark is occluded, as they are trained on non-occluded examples of particular landmarks. This makes them useful as predictors of hand-over-face gesture descriptors. An example of patch expert responses in presence of occlusion can be seen in Figure 5.



Fig. 6. Sample frames from videos that were badly tracked. Note the extreme occlusion and/or head rotations.

4.3. HOG

HOG [Dalal and Triggs 2005] are a popular feature for describing appearance that have been successfully used for pedestrian detection [Dalal and Triggs 2005] and facial landmark detection [Zhu and Ramanan 2012] amongst others.

HOG descriptor counts the number of oriented gradient occurrences in a dense grid of uniformly spaced cells. These occurrences are represented as a histogram for each cell normalised in a larger block area. HOG features capture both appearance and shape information making them suitable for a number of computer vision tasks.

5. EXPERIMENTAL EVALUATION

For our classification tasks, we used the labelled subset of Cam3D described in Section 3.1 to evaluate our approach. It has a total of 365 videos of ~ 2190 s, which contains $\sim 65,700$ frames (~ 6570 data samples, which is one data sample per processing window $w = 10$).

5.1. Methodology

As a pre-processing step, we performed face alignment on all of our videos. Face detection was done using the Zhu and Ramanan [2012] face detector followed by refinement and tracking using the CLNF landmark detector [Baltrušaitis et al. 2013]. After landmark detection, the face was normalised using a similarity transform to account for scaling and in-plane rotation. This led to a 160×120 pixel image, as seen in Figure 1. The output of the facial landmark detection stage was passed to the three feature extraction sub-systems.

The face detector did not manage to initialise in the first frame in all of the videos. To cope with this, we performed backwards tracking alongside forwards tracking from initial detection, leading to more robust landmark detection.

Even with these advanced tracking techniques, our analysis excluded 16 videos, as face detection on them was unsuccessful. Those videos included either extreme head rotation or extreme hand occlusion covering most of the face area that continued throughout the video, thus preventing the tracker from finding a non-occluded frame to recover. Figure 6 shows some examples when face tracking fails.

Space-time features were extracted at the original video frame rate (30 frames per second) using the implementation provided by Laptev et al. [2008]. We removed the features not in the facial region by using the results from the landmark detector. For sparse coding, we used the implementation provided by Mairal et al. [2010] to learn a codebook of size 750 for each training set. The size of the codebook was obtained by trying out different sizes (200, 500, 750) and cross validating across all the videos to obtain the best parameter that produced the minimum data reconstruction error. A user-independent cross validation was utilised for this task. Space-time features were aggregated using max pooling across a window $w = 10$ frames.

Table I. Hand Detection Classification Results (Accuracy and F1 Score) Comparing Uni-Modal and Multi-Modal Feature Fusion. Multi-Modal Fusion of Features Using a Linear SVM Classifier Had the Best Detection Rate (Shown in Bold), Significantly Higher Than a Naïve Baseline

Hand occlusion detection	Majority vote baseline	Uni-modal classification - Linear SVM			Multi-modal classification	
		Likelihood	HOG	STIP	Linear SVM	Non-linear SVM
F1	0.69	0.66	0.82	0.68	0.83	0.80
Acc.	0.56	0.67	0.83	0.56	0.83	0.80

For our task, we extracted HOG features from a similarity normalised 160×120 pixel image of a face. We used 8×8 pixel cells with 18 gradient orientations and block size of 2×2 cells. This led to a 9,576-dimensional HOG descriptor. We reduced its dimensionality using Principal Component Analysis and keeping 90% of the explained variance, leading to 1,035 dimensions vectors per frame. We aggregated the HOG features in a temporal manner by taking the mean value in a window $w = 10$ frames.

As a final feature, we used the landmark alignment likelihoods for each of the 68 landmarks. This was aggregated over a 10-frame window as well by taking its mean.

For classification, our experiments consisted of uni-modal and multi-modal early fusion of extracted features. We used a linear SVM classifier using the Liblinear [Fan et al. 2008] library. We also evaluated SVM classifier with a Radial Basis Function (RBF) to check if this leads to any improvement in performance [Chang and Lin 2011].

The optimal parameters for the SVM were obtained automatically using a leave-one-out cross validation, by holding all videos of one participant out for testing at each iteration. To ensure that our results are generalizable, all experiments were performed in a user-independent approach, as none of the participants in the test set are used for validation or training (both in the classifiers and the dimensionality reduction techniques).

To obtain the ground truth for each classification task, we aggregated the annotations provided by experts (as described in Section 3) for every window $w = 10$ frames. We obtained the ground truth by taking the majority vote across the window of size $w = 10$ frames from the two annotators and assigning the value of the most common label. In case of a tie (disagreement between the labellers) the window w was discarded from further analysis—as this implied that these frames were ambiguous. The total number of frames discarded at this step were less than 10% of the total number of frames in all of the categories.

Besides speeding up the computation time of our approach, the choice of the aggregation window size stemmed from our interest in coding and detecting hand gestures that are semantically higher than frame-level micro-expressions. In other words, we did not expect a change in hand gesture in less than one third of a second.

For all our experiments, we compared our approach performance with chance baseline and a naïve majority vote classifier baseline and evaluated the statistical significance using a Related Samples Friedman’s ANOVA, with follow-up post hoc tests with a Bonferroni correction to p values [Field 2013]. This was chosen as we wanted to perform pairwise comparisons and the data distribution cannot be assumed to follow a normal distribution. The unit of analysis of the significance tests is all the video of one participant (degrees of freedom = 8).

5.2. Hand Occlusion Detection

The first task in our experiments was hand-over-face occlusion detection. The face was considered to be occluded if one or many facial regions are labelled as occluded. For this task, we used a binary classifier to detect if the face is occluded or not. We trained a linear SVM classifier using single modalities and feature-level fusion. Table I shows

Table II. Classification Results (F1 Score and Accuracy) of *Facial Region Occluded* Descriptor Comparing Uni-Modal and Multi-Modal Feature Fusion. Occlusion of Each Face Area Is Treated as a Separate Binary Classification Problem. Multi-Modal Fusion of Features Outperforms Single Modalities in all the Classification Tasks

Facial region		Majority vote	Uni-modal classification - Linear SVM			Multi-modal classification	
			Likelihood	HOG	STIP	Linear SVM	Non-linear SVM
Chin	F1	0.68	0.84	0.68	0.68	0.83	0.84
	Acc.	0.56	0.69	0.85	0.56	0.85	0.87
Lips	F1	0.78	0.88	0.92	0.90	0.94	0.93
	Acc.	0.56	0.82	0.88	0.83	0.90	0.89
Middle face (cheek/nose)	F1	0.73	0.86	0.85	0.87	0.86	0.86
	Acc.	0.61	0.77	0.76	0.77	0.78	0.77

the classification results (accuracy and F1 score) of uni-modal features and multi-modal fusion. We found that the best performance is obtained from the multi-modal linear classifier (Accuracy 0.83, F1 score 0.83), which is higher than a naïve majority vote classifier (Accuracy 0.56, F1 score 0.69) or chance (Accuracy 0.5). To check the significance of the improved classification accuracies, a Related Samples Friedman's ANOVA showed significant difference ($\chi^2(2) = 12.67, p < 0.01$). Further pairwise comparisons showed that our classifier yielded significant improvement over chance baseline ($p < 0.01$) and no significant difference over majority vote ($p = 0.1$).

We also tested the multi-modal fusion in a non-linear SVM, which did not produce better results (Accuracy 0.80, F1 score 0.80). This may be because using a complex kernel has little, if any, impact on the classification performance if we are fusing different features of different representations.

If we look at single modality results, then we notice that the feature that had the highest uni-modal classification results is HOG, which indicates that appearance features can differentiate well between occluded and non-occluded faces, even in the challenging conditions of hand occlusion (see Table I).

5.3. Classification of Hand-Over-Face Gesture Descriptors

After occlusion detection, the second task was to classify hand-over-face gesture descriptors (*hand shape*, *hand action*, and *facial region occluded*). Here, we used a subset of frames where hand occlusion had been identified. We treated each descriptor as a separate classification task. *Hand shape* and *facial region occluded* classifications were performed per window w , while *hand action* classification was done per video.

Facial region occluded descriptor's values are not mutually exclusive, that is, we can have occlusion in more than one face region at any window w . That is why we used three binary classifiers, one for each face region. In each experiment, we used a linear SVM classifier using single features then fused the features in a multi-modal classifier. Table II shows the classification results using these different approaches, highlighting the best obtained result for each classification task.

Having a closer look at the data distribution of different descriptors' values, we found that the data were mostly unbalanced. This is to be expected for this type of problem, because we are analysing gestures in natural expressions with high variance in individual differences, so we do not expect to see all the descriptors' values appearing with the same frequency in all the occlusion videos. This was particularly extreme in the chin region as we had a hand covering the chin in 92% of the occlusion videos. This is not a surprise as the hand would cover the chin in most of the face occlusion gestures as it comes from below the face. To remove the unbalanced effect for the chin classifier, we added more negative samples that were randomly selected from the Cam3D dataset

Table III. Classification Results of *Hand Shape* Descriptor Comparing Uni-Modal and Multi-Modal Feature Fusion as a Four-Class Classification Problem. The Four Classes Are as Follows: Fingers, Closed Hand, Open Hand, and Hands Together. Multi-Modal Fusion of Features Outperforms Single Modalities with an Accuracy That Is Significantly Higher Than the Majority Vote Baseline

Hand shape	Majority vote baseline	Uni-modal classification - Linear SVM			Multi-modal classification	
		Likelihood	HOG	STIP	Linear SVM	Non-linear SVM
Acc.	0.14	0.31	0.35	0.19	0.36	0.36

to the pool of videos used for chin training and classification. A different distribution of the descriptors' values among different participants also presented a challenge in the classification. Since our experiments are user independent, unbalanced distribution of cues presented a challenge to the classifiers.

5.3.1. Facial Region Occluded. Table II shows the classification results for the *facial region occluded* descriptor using the uni-modal and multi-modal classification approaches, highlighting the approach that has the best performance for each task. For chin occlusion detection, multi-modal fusion of features in a non-linear SVM classifier had the best performance (Accuracy 0.87, F1 score 0.84), just slightly higher than the multi-modal linear classification (Accuracy 0.85, F1 score 0.83). For lips occlusion detection, the multi-modal linear SVM classifier had the best performance (Accuracy 0.90, F1 score 0.94). For middle face area occlusion detection (cheeks and nose), the multi-modal linear SVM classifier had the best performance (Accuracy 0.78, F1 score 0.86). This confirms that multi-modal fusion of the feature performed better in all the *facial region occluded* classification tasks.

Related Samples Friedman's ANOVA showed a significant difference between the multimodal fusion classification accuracy, majority vote accuracy, and chance baseline in the three classification tasks (chin detection $\chi^2(2) = 11.56$, $p < 0.01$, lips detection $\chi^2(2) = 12.67$, $p < 0.01$, middle face detection $\chi^2(2) = 12.67$, $p < 0.01$). Further pairwise comparisons showed that multimodal detection results proved to be significantly higher than a naïve chance baseline for the chin, lips, and middle face areas (with $p < 0.01$, $p < 0.05$, and $p < 0.05$, respectively). Multimodal detection results were not significantly higher than the majority vote baseline for the chin, lips, and middle face areas (with $p = 0.06$, $p = 0.10$, and $p = 0.47$, respectively). This is another indication of the unbalanced data.

5.3.2. Hand Shape. Classification of *hand shape* was implemented as a four-class classification problem (one against all), as shape descriptor's values are mutually exclusive per processing window w . The classifier categorised the hand shape as one of four classes: fingers, open hand, closed hand, and hands together (tangled). As shown in Table III, multi-modal fusion of features outperforms single modalities with an accuracy of 0.36 that is higher than the majority vote classification baseline (Accuracy 0.14) and chance baseline (Accuracy 0.25). Related Samples Friedman's ANOVA showed a significant difference among multimodal fusion classification accuracy, majority vote accuracy, and chance baseline ($\chi^2(2) = 12.67$, $p < 0.01$). Further pairwise comparisons showed that my classifier yielded significant improvement over majority vote baseline ($p < 0.05$) but no significant difference with chance ($p = 0.10$).

Classification of the hand shape was challenging due to the similarities between some of the classes. Table IV shows the confusion matrix. The main misclassification instances are "open hand" misclassified as "fingers" and "hands together" misclassified as "closed hand"; this might be because of the similarities in the appearance of these classes. These results indicate that further refinement in the coding scheme of the hand shape descriptor might be needed to improve the classification results.

Table IV. Confusion Matrix Showing Hand Shape Classification Results Using Multimodal Fusion of Features

Predicted \ Ground truth	Ground truth			
	Fingers	Open hand	Closed hand	Hands together
Fingers	715	84	206	32
Open hand	146	26	179	0
Closed hand	596	80	211	122
Hands together	301	0	58	31

Note the misclassification of “open hand” with “fingers” and misclassification of hands together with closed hand indicating the similarities in the appearance of these classes.

Table V. Classification Results (F1 Score and Accuracy) of *Hand Action* Descriptor Comparing Uni-Modal and Multi-Modal Feature Fusion. Classification Performance Remained Very Close to the Majority Vote Baseline, with the Multi-Modal Fusion of Features using a Non-Linear SVM Classifier Having the Best Results

Hand action	Majority vote baseline	Uni-modal - Linear SVM			Uni-modal - non-Linear SVM			Multi-modal	
		Likelihood	HOG	STIP	Likelihood	HOG	STIP	Linear SVM	Non-linear SVM
F1	0.81	0.81	0.81	0.81	0.80	0.82	0.81	0.80	0.83
Acc.	0.70	0.70	0.70	0.70	0.70	0.73	0.70	0.67	0.76

Note that the unbalanced dataset and initial video segmentation criteria in the Cam3D dataset influenced the performance of classifying this descriptor.

5.3.3. Hand Action. For hand action, the data were labelled as one label per video, describing the hand action as static or dynamic in the majority of the video frames. Therefore, we aggregated the features to obtain one feature set per video. Space-time features (STIP) were aggregated using max pooling in the same way described in Section 4.1, and this allowed us to capture the salient features in the sparse codes. For HOG and likelihood features, we calculated means and standard deviations to capture the changes in the features across the video.

We used a binary classification approach to categorise the hand action as dynamic or static. As shown in Table V, SVM linear classification did not perform well on this descriptor, with classification accuracies swinging around the majority vote baseline accuracy, which is 0.7, which is already high due to unbalanced data distribution. Multi-modal classification using a non-linear SVM classifier achieved the highest results (Accuracy 0.76, F1 score 0.83), which are higher than the majority vote (Accuracy 0.70, F1 score 0.81) and chance (Accuracy 0.5). Related Samples Friedman’s ANOVA showed a significant difference among multimodal fusion classification accuracy, majority vote accuracy, and chance baseline ($\chi^2(2) = 9.88, p < 0.01$). Further pairwise comparisons showed that our classifier yielded significant improvement over chance baseline ($p < 0.05$) but no significant improvement over majority vote baseline ($p = 0.9$).

Unbalanced dataset and initial video segmentation criteria in the Cam3D dataset can explain the low increase of the classification results of this descriptor compared to a naïve majority vote classifier, for example: Some video segments have one part of the video with hand motion and the rest without motion, which indeed introduced confusion factor to the classifier. Re-segmenting the videos into shorter segments based on the hand motion would improve the classification accuracy, but we leave this part to future work.

5.4. Discussion

Figure 7 summarises our classification results for hand detection and classification obtained for the six classification tasks. The results display mostly binary classifiers except for hand shape where we employed a four-class classifier, hence the lower classification values. Our multi-modal fusion approach showed a statistically significant improvement over a chance baseline for all of our classification experiments, except for

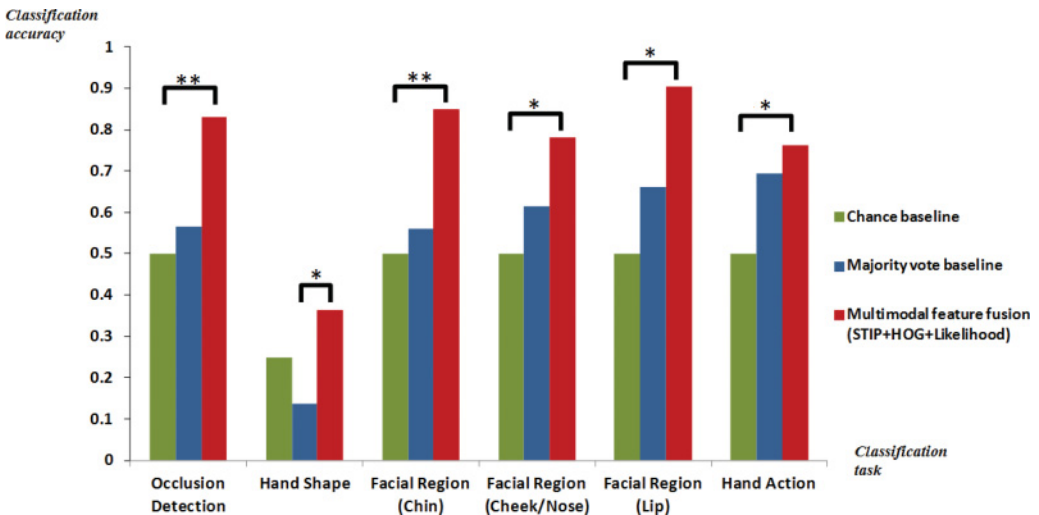


Fig. 7. Classification results summary for all the classification tasks. All are binary classifiers except for (hand shape) where we employ a four-class classifier, hence the lower classification values. Our multi-modal fusion approach showed statistically significant improvement over naive classifier baselines for all of our hand detection and classification tasks (* $p < 0.05$, ** $p < 0.01$).

the hand shape task, where our approach showed statistically significant improvement over majority vote baseline. However, in all the classification tasks our classification accuracy was higher than chance and majority vote baselines.

For the challenging nature and novelty of the gesture classification task, we consider these results satisfactory, considering the nature of the unbalanced dataset with which we are dealing (few training samples for some categories). Unbalanced distribution of the descriptors' values among different participants presented a challenge in the classification as well. Since our experiments are user independent, unbalanced distribution of cues presented a challenge to the classifiers.

6. CONCLUSION AND FUTURE WORK

In this article, we have presented an automatic approach to tackle the challenging problem of detection and classification of hand-over-face gestures. We treat the problem as two tasks: hand occlusion detection, followed by classification of hand gesture cues, namely hand shape, hand action, and facial region occluded. We extract a set of spatial and spatio-temporal features (HOG, facial landmark detection likelihoods, and STIP features). We use feature-specific dimensionality reduction techniques and aggregation over a window of frames to obtain a compact representation of our features. Using a multi-modal classifier of the three features, we can detect hand-over-face occlusions and classify hand shape, hand action, and facial region occluded significantly better than the majority vote and chance baselines. We also demonstrate that multi-modal fusion of the features proved to outperform single modality classification results.

Based on the quantitative analysis presented in this article, there are several future directions for work that we suggest, which include the following.

6.1. Data

More balanced labelled data are needed. Having an automatic detection technique can allow for faster data extraction. For example, our detection approach can be used to quickly scan publicly available video segments to pick the segments that include

hand-over-face gestures for further analysis. Data collection is time consuming, especially that of naturally evoked data. But the future of affective computing will focus on natural rather than acted data [Zhang et al. 2014].

6.2. Refining the Coding Scheme

We have presented a novel coding scheme for hand-over-face gestures, but this scheme has some limitations. For example, as discussed in Section 5.3.2, the proposed coding scheme for hand shape descriptor proved challenging for the proposed automatic classifier due to the similarity of the appearance features of some of the labels, such as “open hand” and “fingers.” Possible future directions can include further refining of this coding scheme, for example, differentiating between separate fingers or between right and left hands. Having access to more data samples can facilitate this step.

6.3. Temporal Machine-Learning Techniques

Considering temporal machine-learning techniques such as Hidden Markov Models or Conditional Random Fields is a possible extension. Taking into account adjacent frame information can improve the performance of hand-over-face gesture detection and classification.

6.4. Multimodal Inference System

Ultimately, our vision is to implement a multimodal affect inference framework that combines facial expressions, head gestures, as well as hand-over-face gestures. Building labelled datasets that include hand-over-face gestures is crucial to be able to achieve such a system. We believe that the work described in this article will open the door for further research in this area.

ACKNOWLEDGMENTS

This work is an extension of the work in Mahmoud et al. [2014], originally published in the proceedings of the International Conference on Multimodal Interaction (ICMI) 2014. The reviewing of this article was managed by the associate editors of the special issue on Highlights of ICMI 2014, Ali Salah, Björn Schuller, Jeff Cohn, Oya Aran, and Louis-Philippe Morency.

REFERENCES

- Hillel Aviezer, Yaacov Trope, and Alexander Todorov. 2012. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* 338, 6111 (2012), 1225–1229.
- Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2013. Constrained local neural fields for robust facial landmark detection in the wild. In *International Conference on Computer Vision (ICCV) Workshops, 300 Faces in-the-Wild Challenge*.
- Felix Barroso and Jason K. Feld. 1986. Self-touching and attentional processes: The role of task difficulty, selection stage, and sex differences. *Journal of Nonverbal Behavior* 10, 1 (1986), 51–64.
- Daniel Bernhardt and Peter Robinson. 2007. Detecting affect from non-stylised body motions. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*.
- Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollar. 2013. Robust face landmark estimation under occlusion. In *International Conference on Computer Vision*.
- Ginevra Castellano, Santiago D. Villalba, and Antonio Camurri. 2007. Recognising human emotions from body movement and gesture dynamics. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*.
- Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* (2011).
- David Cristinacce and Tim Cootes. 2006. Feature detection and tracking with constrained local models. In *British Machine Vision Conference (BMVC)*.
- Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *The International Conference on Computer Vision and Pattern Recognition (CVPR)*.

- Beatrice de Gelder. 2009. Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3475–3484.
- Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. 2005. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance*.
- Cao Dong, Ming Leu, and Zhaozheng Yin. 2015. American sign language alphabet recognition using microsoft kinect. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 44–52.
- Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research* 9 (2008), 1871–1874.
- Andy Field. 2013. *Discovering Statistics using IBM SPSS Statistics*. Sage, Thousand Oaks, CA.
- Norbert Freedman. 1977. Hands, words, and mind: On the structuralization of body movements during discourse and the capacity for verbal representation. In *Communicative Structures and Psychic Structures*. Springer, Berlin, 109–132.
- Matilde Gonzalez, Christophe Collet, and Rémi Dubot. 2012. Head tracking and hand segmentation during hand over face occlusion in sign language. In *Trends and Topics in Computer Vision*. Springer, Berlin, 234–243.
- Joseph F. Grafsgaard, Robert M. Fulton, Kristy Elizabeth Boyer, Eric N. Wiebe, and James C. Lester. 2012. Multimodal analysis of the implicit affective channel in computer-mediated textual communication. In *International Conference on Multimodal Interaction (ICMI)*.
- Martin Grunwald, Thomas Weiss, Stephanie Mueller, and Lysann Rall. 2014. EEG changes caused by spontaneous facial self-touch may represent emotion regulating processes and working memory maintenance. *Brain Research* 1557 (2014), 111–126.
- Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* (2007).
- Kazuhiro Hotta. 2004. A robust face detector under partial occlusion. In *The International Conference on Image Processing (ICIP)*.
- Horace H. S. Ip, Sam C. S. Chan, and Maria S. W. Lam. 1998. HACS: Hand action coding system for anatomy-based synthesis of hand gestures. In *IEEE International Conference on Systems, Man, and Cybernetics*, Vol. 2. IEEE, Los Alamitos, CA, 1207–1212.
- C. Keskin, F. Kirac, Y. E. Kara, and L. Akarun. 2011. Real time hand pose estimation using depth sensors. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, Los Alamitos, CA, 1228–1234.
- Andrea Kleinsmith and Nadia Bianchi-Berthouze. 2013. Affective body expression perception and recognition: A survey. *IEEE Transaction on Affective Computing* 4, 1 (2013), 15–33.
- Klaus Krippendorff. 2004. *Content Analysis: An Introduction to its Methodology*. Sage, Thousand Oaks, CA.
- A. Kurakin, Z. Zhang, and Z. Liu. 2012. A real time system for dynamic hand gesture recognition with a depth sensor. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, Los Alamitos, CA, 1975–1979.
- Ivan Laptev. 2005. On space-time interest points. *International Journal of Computer Vision* 64, 2–3 (2005), 107–123.
- Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. 2008. Learning realistic human actions from movies. In *The International Conference on Computer Vision and Pattern Recognition (CVPR)*.
- H. Lausberg and H. Sloetjes. 2009. Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods* 41, 3 (2009), 841–849. <http://www.lat-mpi.eu/tools/elan/>.
- Marwa Mahmoud, Tadas Baltrušaitis, Peter Robinson, and Laurel D. Riek. 2011. 3D corpus of spontaneous complex mental states. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*.
- Marwa Mahmoud, Rana El-Kaliouby, and Amr Goneid. 2009. Towards communicative face occlusions: Machine detection of hand-over-face gestures. In *Image Analysis and Recognition*. Springer, Berlin, 481–490.
- Marwa Mahmoud and Peter Robinson. 2011. Interpreting hand-over-face gestures. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*.
- Marwa M. Mahmoud, Tadas Baltrušaitis, and Peter Robinson. 2014. Automatic detection of naturalistic hand-over-face gesture descriptors. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, Istanbul, Turkey, 319–326.

- Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. 2010. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research* (2010).
- Georg Poier, Konstantinos Roditakis, Samuel Schulter, Damien Michel, Horst Bischof, and Antonis A. Argyros. 2015. Hybrid one-shot 3d hand pose estimation by exploiting uncertainties. In *British Machine Vision Conference (BMVC)*.
- Ronald Poppe. 2010. A survey on vision-based human action recognition. *Image and Vision Computing* 28, 6 (2010), 976–990.
- Zhou Ren, Jingjing Meng, Junsong Yuan, and Zhengyou Zhang. 2011. Robust hand gesture recognition with kinect sensor. In *Proceedings of the 19th ACM International Conference on Multimedia*. ACM, New York, NY, 759–760.
- Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. 2011. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision* 91, 2 (2011), 200–215.
- Paul Smith, Niels da Vitoria Lobo, and Mubarak Shah. 2007. Resolving hand over face occlusion. *Image and Vision Computing* 25, 9 (2007), 1432–1448.
- Yale Song, Louis-Philippe Morency, and Randall Davis. 2013. Learning a sparse codebook of facial and body microexpressions for emotion recognition. In *International Conference on Multimodal Interaction (ICMI)*.
- Jesus Suarez and Robin R. Murphy. 2012. Hand gesture recognition with depth images: A review. In *IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 411–417.
- Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, Cordelia Schmid, and others. 2009. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference (BMVC)*.
- Cheoljong Yang, Yujeong Jang, Jounghoon Beh, David Han, and Hanseok Ko. 2012. Gesture recognition using depth-based hand tracking for contactless controller application. In *IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 297–298.
- Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. 2009. Linear spatial pyramid matching using sparse coding for image classification. In *The International Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Xiang Yu, Fei Yang, Junzhou Huang, and Dimitris N. Metaxas. 2013. Explicit occlusion detection based deformable fitting for facial landmark localization. In *10th IEEE Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–6.
- Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran. 2014. Facial expression recognition experiments with data from television broadcasts and the world wide web. *Image and Vision Computing* 32, 2 (2014), 107–119.
- Xiangxin Zhu and Deva Ramanan. 2012. Face detection, pose estimation, and landmark localization in the wild. In *The International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Received June 2015; revised October 2015; accepted February 2016