Automatic Detection of Naturalistic Hand-over-Face Gesture Descriptors

Marwa Mahmoud, Tadas Baltrušaitis and Peter Robinson
University of Cambridge, Computer Laboratory, UK
{marwa.mahmoud, tadas.baltrusaitis, peter.robinson}@cl.cam.ac.uk

ABSTRACT

One of the main factors that limit the accuracy of facial analysis systems is hand occlusion. As the face becomes occluded, facial features are either lost, corrupted or erroneously detected. Hand-over-face occlusions are considered not only very common but also very challenging to handle. Moreover, there is empirical evidence that some of these hand-over-face gestures serve as cues for recognition of cognitive mental states. In this paper, 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 higher than a naïve baseline. To our knowledge, this work is the first attempt to automatically detect and classify hand-over-face gestures in natural expressions.

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: Video analysis

Keywords

Hand-over-face occlusions; Hand cues; Facial landmarks; Histograms of oriented gradient; Space-time interest points

1. INTRODUCTION

Over the past few years, there has been an increased 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 either 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 occlusion can significantly improve facial landmark detection and facial expression inference systems.

Moreover, hand-over-face occlusions are not just noise that needs to be removed. Recent studies show that body movements and gestures are significant visual cues that complement facial expressions [8] and they can be utilised in automatic detection of human internal states [2, 4]. Specifically, hand-over-face gestures can serve as an additional valuable channel for multi-modal affect inference [22]. 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 automatically quantify and detect these gestures, instead of the common practice of manual coding. To date, there is no available automatic detection system that serves these purposes.

In this paper, 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 paper are:

1. Proposing a mutli-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

ICMI'14, November 12–16, 2014, Istanbul, Turkey

ACM 978-1-4503-2885-2/14/11 $15.00.
http://dx.doi.org/10.1145/2663204.2663258.

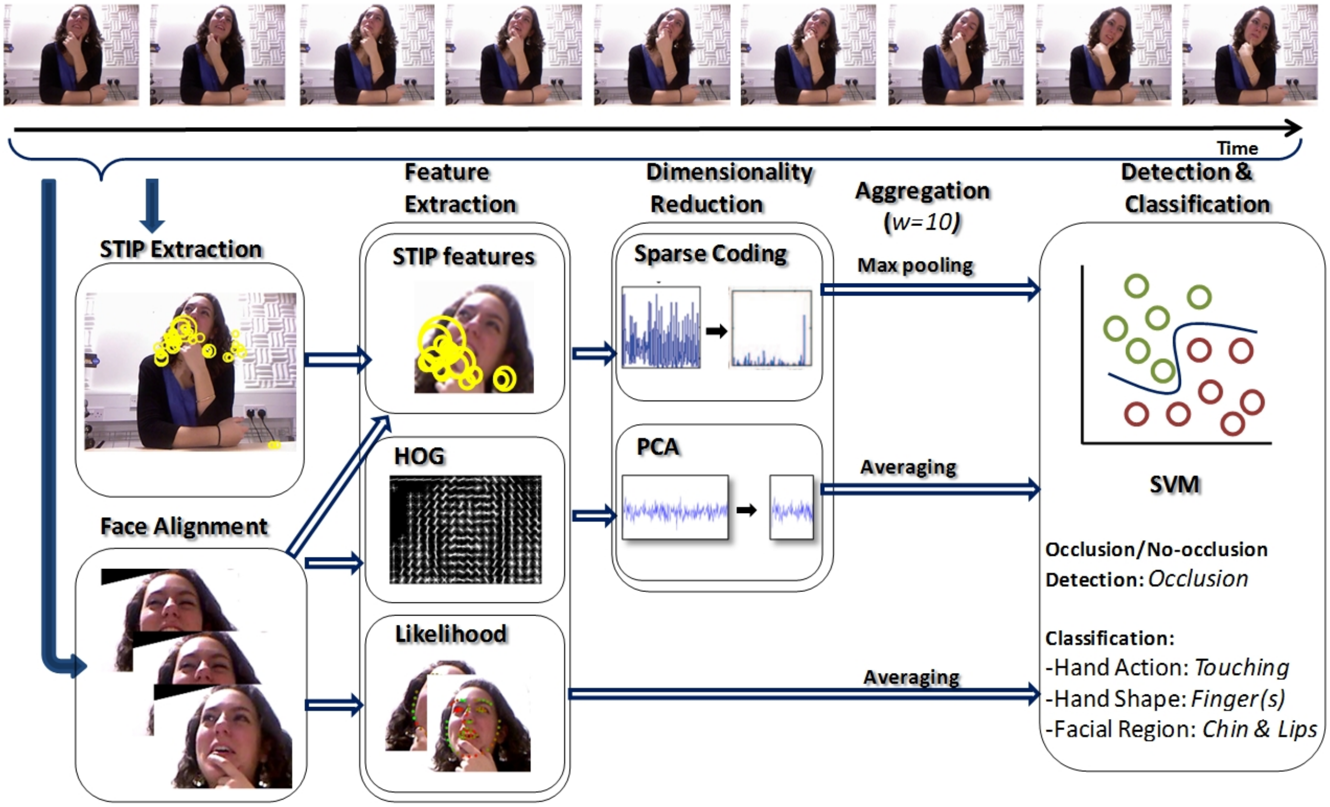

Figure 1: Overview Diagram shows the main steps in our approach.

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. [30] and Burgos-Artizzu et al. [3]. 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. 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 [15], that proposes a method for more robust face detection in presence of partial occlusion by using SVM with local kernels. However, none of the above mentioned authors distinguish between types of occlusion and make no special analysis of hand-over-face occlusions.

There have been a few attempts [26, 21] to detect the hand when it occludes the face. Gonzalez et al. [12] use colour and edge information to track and segment the hand during hand over face occlusion in sign language. Grafsgaard et al. [13] use 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 different classification of hand-over-face gesture descriptors.

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. Since most of the work on affect analysis focuses on the face, most of the publicly available natural datasets also focus on faces with limited or no occlusion. Since we are interested in the temporal aspect of the hand gesture as well, still photograph corpora did not satisfy our criteria. The publicly available Cam3D [20] 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.

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 9 participants with mean duration of the video being 6 seconds. 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 chosen no-occlusion videos were selected containing the same 9 participants while keeping the number of samples per 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. The goal was to code hand gestures in terms of certain descriptors that can describe the gesture. Inspired by the coding schema provided by Mahmoud et al. [22], we coded the gestures in terms of *hand shape*, *hand action* and *facial region occluded*.

Labelling was carried out using Elan video annotation tool [19]. 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: 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, i.e. one label is permitted per frame. Labels are: 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: 1) **Forehead**. 2)**Eye(s)**. 3)**Nose**. 4)**Cheek(s)**. 5)**Lips**. 6)**Chin**. 7)**Hair/ear**.

3.3 Coding assessment & 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 [16], which is widely used for this type of coding assessment because of its independence of the number of assessors and its robustness against imperfect data [14]. 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 facial regions: eyes, nose and hair/ear. When we explored the reason of the disagreement in these categories, this was mostly because of the very few samples available of these categories in the dataset, for example: eyes, forehead and hair/ear regions had only 25, 100 and 10 frame samples respectively, i.e. less than 0.2% of the total number of frames in total. We decided to exclude these categories (mostly upper face area) in the machine learning step, as it was unfair to try to automatically learn and classify these categories when the human annotators failed to agree.

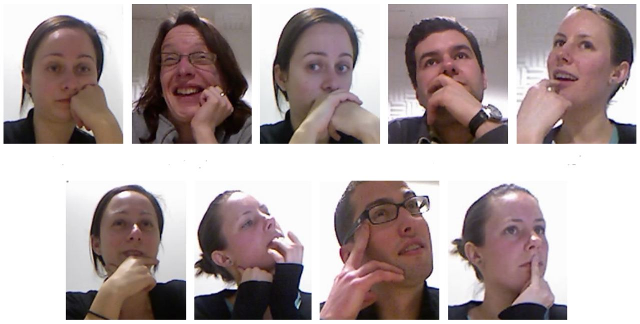

Figure 2: Sample frames from videos in the dataset Cam3D showing examples of face touches present in the dataset [20]. Note the challenging - close to natural - recording settings like inconsistent lighting conditions and strong head rotations.

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 in one group representing non-static hand gesture, i.e. any type of motion.

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) and facial landmark alignment likelihood. Moreover, having the detection of hand action in mind, we also extract Space Time Interest Points (STIP) 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 [17, 18, 9] have become popular motion descriptors for action recognition [24]. Recently, they have been used by Song et al. [27] 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. Our methodology for space time interest points feature extraction and representation is based on the approach proposed by Song et al. [27].

Space Time Interest Points (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: spatio-temporal interest point (STIP) detection followed by feature extraction. Wang et al. [28] reports that using the Harris3D interest point detector followed by a combination of the Histograms of Oriented Gradient (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 sparse-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 that only few salient features are recovered, i.e. 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, i.e. the number of codebook entries is larger than the dimension of input signal so that only a small number of codebook entries are used to represent the input signal. Given an input signal $\mathbf{x} \in \mathbb{R}^{\mathbf{N}}$ and over-complete codebook $\mathbf{D} \in \mathbb{R}^{\mathbb{N}\times\mathbb{K}}$, $\mathbf{K} \gg \mathbf{N}$, we find a sparse signal $\alpha \in \mathbb{R}^{\mathbf{K}}$ that minimises the reconstruction error,

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

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 that 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, i.e. the extracted space-time features $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$

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

The optimisation problem is convex in $\mathbf{D}$ with $\mathbf{A} = [\alpha_1, \alpha_2, ... , \alpha_M]$ fixed and in $\mathbf{A}$ with $\mathbf{D}$ fixed, but not in both at the same time [23]. Thus, it can be solved using online learning [23] 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 better representation that is invariant to image transformations and noise [29, 27]. The max-pooling operation is defined as:

$$\mathbf{z} = [\max_{i=1...M_v} |\alpha_{i,1}|, \max_{i=1...M_v} |\alpha_{i,2}|, ..., \max_{i=1...M_v} |\alpha_{i,\mathbf{K}}|], \tag{3}$$

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.

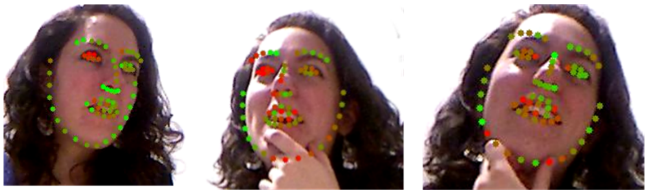

Figure 3: An example of patch expert responses in presence of occlusion. Green shows high likelihood values, while red means low likelihoods.

We employ a Constrained Local Neural Field (CLNF) [1] 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) [6], that uses more advanced patch experts and optimisation function. We use the publicly available CLNF implementation [1].

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 3D rotation matrix); 2D translation $\mathbf{t}$; a vector describing non-rigid variation of shape $\mathbf{p}$. The point distribution model (PDM) is:

$$\mathbf{x}_i = s \cdot \mathbf{R}(\overline{\mathbf{x}}_\mathbf{i} + \mathbf{\Phi}_i\mathbf{p}) + \mathbf{t}. \tag{4}$$

Here $\mathbf{x}_i = (x, y)$ denotes the 2D location of the i^{th} feature point in an image, $\overline{\mathbf{x}}_i = (X, Y, Z)$ is the mean value of the i^{th} element of the PDM in the 3D reference frame, and the vector $\mathbf{\Phi}_i$ is the i^{th} eigenvector obtained from the training set that describes the linear variations of non-rigid shape of this feature point, and the vector $\mathbf{\Psi}_i$ 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 $\mathbf{p}$:

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

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 4.2.1).

We employ a common two step CLM fitting strategy [6, 25]; 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 [1].

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 [31] 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; 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 [1], and is too slow to be used for video analysis.

4.2.1 *Likelihood*

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 [1], 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 3.

4.3 HOG

Histograms of Oriented Gradients (HOG) [7] are a popular feature for describing appearance that has been successfully used for pedestrian detection [7], and facial landmark detection [31] 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 seconds, which contains ~ 65700 frames (~ 6570 data samples - 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 Zhu and Ramanan's [31] face detector followed by refinement and tracking using CLNF landmark detector [1]. 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 4 shows some examples).

Space time features were extracted at the original video frame rate (30 frames per second) using the implementation provided by Laptev *et al.* [18]. 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. [23] 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.

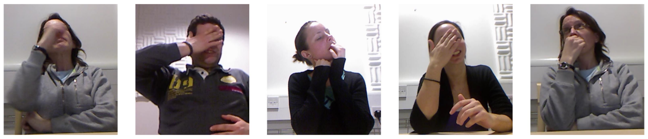

Figure 4: Sample frames from videos that were badly tracked. Note the extreme occlusion and/or head rotations.

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 9576 dimensional HOG descriptor. We reduced its dimensionality using Principal Component Analysis and keeping 90% of explained variance, leading to 1035 dimensions vector 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 Liblinear [10] library. We also evaluated an RBF kernel SVM classifier to check if this leads to any improvement in performance [5].

The optimal parameters for the SVM were automatically obtained 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 a fol-

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

Table 1: Hand detection classification results 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.

low up post-hoc tests with a Bonferroni correction to p values [11]. This was chosen as we wanted to perform pair-wise comparisons and the data distribution cannot be assumed to follow a normal distribution.

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 1 shows 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 results, statistical tests showed that our classifier yielded significant improvement over chance ($p = 0.001$).

We also tested the muti-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, 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.

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*). 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, i.e. 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 2 shows the classification results using these different approaches, highlighting the best obtained result for each classification task.

Taking a closer look at the data distribution of different descriptors' values, we found that the data was mostly unbalanced. This is expected for this type of problems 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 hand covering 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 Cam3D dataset to the pool of videos used for chin training and classification. 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 2 shows the classification results for *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 mutli-modal linear classification (Accuracy 0.85, F1 score 0.83). For lips occlusion detection, 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), 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. Our detection results proved to be statistically significantly higher than a naïve chance baseline for the chin, lips and middle face areas (with p=0.003, p=0.001 and p=0.014 respectively).

5.3.2 Hand Shape

Classification of *hand shape* was implemented as a 4 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 3, Multi-modal fusion of features outperforms single modalities with an accuracy of 0.36 that is significantly higher than the majority vote classification baseline (Accuracy 0.14) ($p = 0.001$) and chance baseline (Accuracy 0.25).

5.3.3 Hand Action

For Hand action, the data was 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, 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.

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

Table 2: Classification results 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.

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

Table 3: Classification results of *hand shape* descriptor comparing uni-modal and multi-modal feature fusion as a 4 class classification problem. The four classes are: 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.

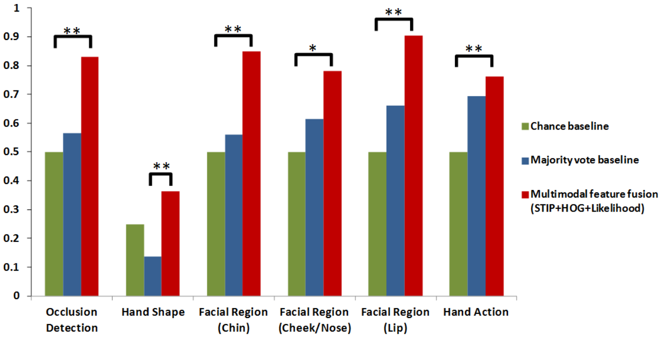

Figure 5: Classification results summary for all the classification tasks. All are binary classifiers except for *(*hand shape) where we employ a 4 class classifier, hence the lower classification values. Our multi-modal fusion approach showed statistically significant improvement over naive classifier baselines for all of our hand detection and classification tasks. (* p<0.05, p<0.01)**

We used a binary classification approach to categorise the hand action as dynamic or static. As shown in Table 4, 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 is higher than the majority vote and significantly higher than chance (p=0.007) . Unbalanced dataset and initial video segmentation criteria in 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 5 summarises our classification results for hand detection and classification obtained for for the six classification tasks. The results display mostly binary classifiers except for hand shape where we employed a 4 class classifier, hence the lower classification values. Our multi-modal fusion approach showed a statistically significant improvement over a naive classifier for all of our classification experiments.

For the challenging nature and novelty of the gesture classification task, we consider these results satisfactory, considering the nature of the unbalanced dataset we are dealing with (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 paper, we presented the first 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, then classification of hand gesture cues, namely - *hand shape*, *hand action* and *facial region occluded* . We extract a set of spatial and spatio-temporal features (Histograms of Oriented Gradients (HOG), facial landmark detection likelihood, and space-time interest points (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 muti-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 mutli-modal fusion of the features proved to outperform single modality classification results.

We believe that adding more temporal features and improving the segmentation of the videos would improve the *hand action* detection results but we are leaving this to fu-

Hand action	Majority	Uni-modal - Linear SVM			Uni-modal - non-Linear SVM			**Multi-modal classification**	
classification	vote	Likelihood	HOG	STIP	Likelihood	HOG	STIP	Linear SVM	**Non-linear SVM**
F1 score	0.81	0.81	0.81	0.81	0.80	0.82	0.81	0.80	**0.83**
Accuracy	0.70	0.70	0.70	0.70	0.70	0.73	0.70	0.67	**0.76**

Table 4: Classification results 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. Note that the unbalanced dataset and initial video segmentation criteria in Cam3D dataset influenced the performance of classifying this descriptor.

ture work. Future work also includes testing our multimodal features using more complex classifiers that incorporate temporal features. The coding schema can also be improved to include more hand articulations.

7. ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 289021.

8. REFERENCES

[1] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *ICCV Workshops, 300 Faces in-the-Wild Challenge*. 2013.

[2] D. Bernhardt and P. Robinson. Detecting affect from non-stylised body motions. In *ACII*. 2007.

[3] X. P. Burgos-Artizzu, P. Perona, and P. Dollar. Robust face landmark estimation under occlusion. In *International Conference on Computer Vision*, 2013.

[4] G. Castellano, S. D. Villalba, and A. Camurri. Recognising human emotions from body movement and gesture dynamics. In *ACII*. 2007.

[5] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011.

[6] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, 2006.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[8] B. de Gelder. Why bodies? twelve reasons for including bodily expressions in affective neuroscience. *Phil. Trans. of the Royal Society B*, 2009.

[9] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005.

[10] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9, 2008.

[11] A. Field. *Discovering statistics using IBM SPSS statistics*. Sage, 2013.

[12] M. Gonzalez, C. Collet, and R. Dubot. Head tracking and hand segmentation during hand over face occlusion in sign language. In *Trends and Topics in Computer Vision*, pages 234–243. Springer, 2012.

[13] J. F. Grafsgaard, R. M. Fulton, K. E. Boyer, E. N. Wiebe, and J. C. Lester. Multimodal analysis of the implicit affective channel in computer-mediated textual communication. In *ICMI*, 2012.

[14] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 2007.

[15] K. Hotta. A robust face detector under partial occlusion. In *ICIP*, 2004.

[16] K. Krippendorff. *Content analysis: An introduction to its methodology*. Sage, 2004.

[17] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[18] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[19] H. Lausberg and H. Sloetjes. Coding gestural behavior with the NEUROGES-ELAN system. *Behavior research methods*, 2009.

[20] M. Mahmoud, T. Baltrušaitis, P. Robinson, and L. D. Riek. 3D corpus of spontaneous complex mental states. In *ACII*. 2011.

[21] M. Mahmoud, R. El-Kaliouby, and A. Goneid. Towards communicative face occlusions: machine detection of hand-over-face gestures. In *Image Analysis and Recognition*, pages 481–490. Springer, 2009.

[22] M. Mahmoud and P. Robinson. Interpreting hand-over-face gestures. In *ACII*. 2011.

[23] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 2010.

[24] R. Poppe. A survey on vision-based human action recognition. *IVC*, 2010.

[25] J. Saragih, S. Lucey, and J. Cohn. Deformable Model Fitting by Regularized Landmark Mean-Shift. *IJCV*, 2011.

[26] P. Smith, N. da Vitoria Lobo, and M. Shah. Resolving hand over face occlusion. *IVC*, 2007.

[27] Y. Song, L.-P. Morency, and R. Davis. Learning a sparse codebook of facial and body microexpressions for emotion recognition. In *ICMI*, 2013.

[28] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, et al. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[29] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[30] X. Yu, F. Yang, J. Huang, and D. N. Metaxas. Explicit Occlusion Detection based Deformable Fitting for Facial Landmark Localization. In *FG*, 2013.

[31] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.