# Analyzing Trajectories on Grassmann Manifold for Early Emotion Detection from Depth Videos

Taleb Alashkar, Boulbaba Ben Amor, Stefano Berretti, Mohamed Daoudi

# Analyzing Trajectories on Grassmann Manifold for Early Emotion Detection from Depth Videos

Taleb Alashkar[1], Boulbaba Ben Amor[1], Stefano Berretti[2], and Mohamed Daoudi[1]
[1]Institut Mines-Télécom/Télécom Lille ; CRIStAL (UMR CNRS 9189), France.
[2]Department of Information Engineering, University of Florence, Italy.

***Abstract*— This paper proposes a new framework for online detection of spontaneous emotions from low-resolution depth sequences of the upper part of the body. To face the challenges of this scenario, depth videos are decomposed into subsequences, each modeled as a linear subspace, which in turn is represented as a point on a Grassmann manifold. Modeling the temporal evolution of distances between subsequences of the underlying manifold as a one-dimensional signature, termed *Geometric Motion History*, permits us to encompass the temporal signature into an early detection framework using Structured Output SVM, thus enabling online emotion detection. Results obtained on the publicly available Cam3D Kinect database validate the proposed solution, also demonstrating that the upper body, instead of the face alone, can improve the performance of emotion detection.**

## I. INTRODUCTION

With the widespread diffusion of devices endowed with onboard cameras (e.g., hand-held devices, entertainment consoles, personal computers, surveillance and monitoring sensors) there is now an increasing interest in performing online detection and recognition of expressions and emotional states. This has many potential applications, such as human-computer interaction, gaming, augmented and virtual reality, drivers fatigue detection, etc. The first studies on facial expressions focused on 2D imagery [1], but in these days, it is a shared conviction that facial expressions are determined by a dynamic process, which can be better interpreted through the analysis of video sequences, rather than the analysis of still images. More recently, some approaches analyzed expressions as spatio-temporal deformations of 3D faces caused by the action of facial muscles. In this case, the facial expressions can be studied comprehensively by analyzing the temporal dynamics of 3D face scans (3D plus time is often regarded as 4D data or 3D dynamic data). From this perspective, the relative immunity of 3D scans to lighting conditions and pose variations gives support to the use of 3D and 4D data. Motivated by these considerations, there has been a progressive shift from 2D to 3D in performing facial shape analysis for facial expression recognition [2], [3]. This trend has been strengthened further by the introduction of inexpensive acquisition devices accessible to a large number of users, such as the Kinect-like cameras that provide fast, albeit low-resolution, streams of 3D data. This opened the way to new opportunities and challenges for facial expression and emotion recognition.

The work of psychologists, which describes human affects in terms of *discrete categories*, has largely influenced the way which most of the approaches use to classify facial expressions. The most popular example of this categorical description is given by the six prototypical (basic) emotion categories, which include *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*.This description was specially supported by the cross-cultural studies conducted by Ekman [4], indicating that humans perceive certain basic emotions with respect to facial expressions in the same way, regardless of culture.

However, this discrete list of emotions fails to describe the range of emotions that occur in natural face-to-face communication. An alternative to the categorical description of human affect is the *dimensional description* [5], in which an affective state is characterized in terms of a small number of latent dimensions, rather than a small number of discrete emotion categories. In particular, the *evaluation* and *activation* dimensions are expected to reflect the main aspects of emotions: The *evaluation* dimension measures how a human feels, from positive to negative; The *activation* dimension measures whether humans are more or less likely to take an action under the emotional state, from active to passive [6]. The dimensional description of emotions is shown in Fig. 1, using the *Arousal-Valence* chart.

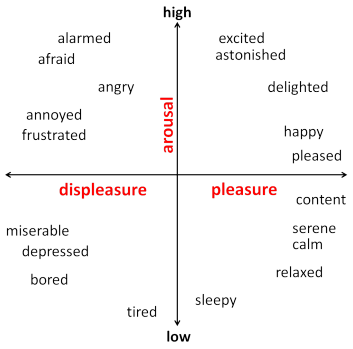


Fig. 1. Dimensional Arousal-Valence description of emotions.

In addition to the limitations posed by the rigid categorical classification, most of the current solutions for facial expression recognition from 3D dynamic sequences are evaluated in constrained scenarios [7], which include high-resolution posed datasets acquired in rigid settings (see for example the BU-4DFE dataset [2]). Instead, the recognition of spontaneous facial expressions is a more challenging problem that recently attracted high interest (see for example

the works in [8] and [9]). The effect of low-resolution noisy acquisitions on expression recognition has been not considered in these studies.

However, the majority of methods yet propose expression classification based on the observation of the entire 3D dynamic sequence (i.e., a decision is taken once the full sequence is observed). No emphasis is placed on the responsiveness, that is on the capability to produce a correct classification just from a partial observation, as short as possible, of the sequence. Indeed, the trade-off between the accuracy and observation size for rapid recognition is an important topic in a wide spectrum of real applications. Schindler and Van Gool [10], first investigated this aspect by evaluating how many frames were required to enable action classification in RGB videos. Su et al. [11] presented a high-frame rate 3D facial expressions recognition system, based on an early AdaBoost classifier, but the test dataset was limited to few subjects and the facial expressions were posed, with a very high temporal resolution. In [12], Su and Sato proposed an early recognition framework based on RankBoost with application to facial expression recognition. More recently, Hoai and De la Torre [13] proposed a learning formulation for early event detection. Their maximum-margin framework is devised for training temporal event detectors capable of recognizing partial events, thus enabling early detection with minimal latency. Their method extends the Structured Output SVM to accommodate sequential data.

A further aspect that has been rarely considered in the literature is the relevance of the body language, in addition to facial expressions, for transmitting emotions [14]. In particular, several studies from different domains agreed that combining the face and body expressions can improve the recognition of emotional states [15], [16]. The joint consideration of these aspects is now fostered by the advancement in acquisition technologies, which allows capturing of 3D depth data of the body as well as the face. This approach has been recently used to improve the understanding of the human-machine interaction [17].

Based on the above premises, this work proposes an online detection approach, capable of recognizing emotions as early as possible, according to a dimensional description. The proposed solution is applied to a challenging scenario, where depth sequences of the upper part of the body are acquired with a low-resolution sensor. Besides, the spontaneous emotions depend on the facial expressions, as well as the dynamics of the upper part of the body.

The rest of the paper is organized as follows: Sect. II outlines the main ideas and contributions of the proposed approach; In Sect. III, the proposed representation of a video sequence as a set of points on a Grassmann manifold is presented. In Sect. IV, the 3D dynamic sequence representation is adapted to an early event-detector framework which permits emotion identification as early as possible. The potential of the proposed solution is showcased in Sect. V, by reporting results from the Cam3D Kinect database. Finally, conclusions and future work are discussed in Sect. VI.

## II. METHOD OVERVIEW AND CONTRIBUTIONS

In this paper, we target an online emotion detection approach capable of working on depth sequences of the upper part of the body acquired using cost-effective cameras. To this end, we utilize two growing, but disparate ideas in computer vision: dynamic data analysis using tools from differential geometry; and early event detection using an adaptation of the Structured Output SVM (SOSVM) to sequential data. The depth frames of a given sequence are first grouped into subsequences of a predefined number of adjacent frames. Each group is regarded as a linear subspace (i.e., span of an orthonormal basis, represented by a matrix). These subspaces are naturally viewed as elements of a Grassmann manifold, which spans linear subspaces of same dimension. Then, geometric tools related to the underlying manifold are used to evaluate the differences between points representing different subsequences (the velocity vector computed between two points results into a geodesic distance on the Grassmannian manifold). Fig. 2 shows the idea of mapping subsequences of depth frames to the Grassmann manifold. The positions of points corresponding to successive subsequences capture the temporal evolution (dynamics) of the upper part of the body, as a trajectory on the manifold. We then consider the temporal evolution of the distances between points across the trajectory as a one-dimensional feature vector called *Geometric Motion History* descriptor. The extracted temporal signature is presented to an early-event detection framework, similar to that proposed in [13], thus enabling online event detection as early as possible.

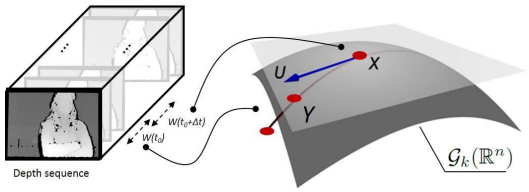


Fig. 2. Dynamic data representation on the Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$.

In summary, the main contributions of this work are:

- A method to represent a depth video through a linear subspace analysis that maps subsequences of frames to points on a Grassmann manifold. In this way, the set of subspaces forms a trajectory on the manifold;
- A method for early detection of emotions, represented according to a dimensional description, uses the dynamics of the upper part of the body. In so doing, we also report a clear benefit in emotion detection using the upper part of the body, rather than the face alone.

We also emphasize that the proposed framework is the first one, to our knowledge, capable of addressing early detection of spontaneous emotions in a complex scenario that includes:

- Depth sequences of the upper part of the body acquired with a cost-effective Kinect camera;
- Spontaneous emotions acquired without a rigid protocol (i.e., no assumption on the time when the emotion occurs in the sequence);

– Emotions related not only to the temporal dynamics of the 3D face deformations, but also to the posture and movement of the head and of the upper part of the body, including shoulders and arms.

## III. TRAJECTORIES ON STIEFEL AND GRASSMANN MANIFOLDS

An important advantage of using dynamic depth flows is that it is relatively easy to isolate and track the human body in the observed scene. Moreover, the depth maps are independent of the appearance and the illumination changes and provide a more complete shape representation of the human body. Despite the limitations due to noise and low-resolution data, adding the temporal dimension to 3D acquisitions is motivated because the body is a deformable 3D surface changing over time. Thus, using the temporal component can be useful to improve the recognition. This shift to the analysis of dynamic data (videos) is now well established in the 2D domain. To overcome the above-mentioned limitations, we propose to use matrix manifold representations and derive geometric tools to analyze dynamic 3D data. Formally, after isolating the human body from the background in the depth images and normalizing the number of pixels to $n$, as a pre-processing step, a window of successive frames $W(t_0)$ (being $t_0$ the starting time of the subsequence) is mapped to a *Stiefel manifold* using $k$-SVD (Singular Value Decomposition) method. The Stiefel manifold $\mathcal{V}_k(\mathbb{R}^n)$ is the set of $n$-by-$k$ tall-skinny orthonormal matrices. The same procedure is applied to the window of frames $W(t_0 + i\Delta t)$ seen at $t_0+i\Delta t$, where $i \in \{0, \ldots, T\}$. As a result, the original depth video is mapped onto the manifold and viewed as a trajectory or a curve (see Fig. 2). The problem of such representation in $\mathcal{V}_k(\mathbb{R}^n)$ is that two matrices $M$ and $M'$ can span the same subspace.

Unlike Stiefel manifold, points on Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$ are equivalence classes of matrices in $\mathcal{V}_k(\mathbb{R}^n)$, where two matrices are equivalent if their columns span the same $k$-dimensional subspace. In other words, $\mathcal{G}_k(\mathbb{R}^n)$ is a quotient space of $\mathcal{V}_k$ ($\mathcal{G}_k(\mathbb{R}^n) = \mathcal{V}_k(\mathbb{R}^n)/O(k)$, where $O(k)$ is the orthogonal group of dimension $k$). Putting it differently, $\mathcal{G}_k(\mathbb{R}^n)$ is the set of all orbits of $\mathcal{V}_k(\mathbb{R}^n)$ under the group action $O(k)$.

Now, to quantify the distance between points on the Stiefel or the Grassmann manifolds, appropriate metrics must be defined. Let us consider arbitrary elements $Y_1, Y_2 \in \mathcal{V}_k(\mathbb{R}^n)$ and $\mathcal{Y}_1 = Span(Y_1)$, $\mathcal{Y}_2 = Span(Y_2) \in \mathcal{G}_k(\mathbb{R}^n)$; the following metrics can be defined as follow,

– **Metric on Stiefel manifold**: The *Frobenius metric* defined by $d_{\mathcal{V}}(Y_1, Y_2) = \|Y_1 - Y_2\|_F$, where $\|\ .\ \|_F$ is the standard Frobenius norm $\|A\|_F = \sqrt{tr(AA^t)}$;
– **Metric on Grassmann manifold**: Golub and Loan [18] introduced an intuitive and computationally efficient way of defining the distance between two linear subspaces using the principal angles. In fact, there is a set of principal angles $\Theta = [\theta_1, \ldots, \theta_k]$ between $\mathcal{Y}_1$ and $\mathcal{Y}_2$, defined as follows:

$$\theta_i = \cos^{-1}\left(\max_{u_i \in \mathcal{Y}_1} \max_{v_i \in \mathcal{Y}_2} \langle u_i^t, v_i \rangle\right), \quad (1)$$

where $u$ and $v$ are the vectors of the basis spanning, respectively, the subspaces $\mathcal{Y}_1$ and $\mathcal{Y}_2$, subject to the additional constraints $\langle u^t, u \rangle = \langle v^t, v \rangle = 1$, and $\langle u^t, v \rangle = \langle v^t, u \rangle = 0$, where $\langle , \rangle$ denotes the inner product in $\mathbb{R}^n$. Based on the definition of the principal angles, the geodesic distance between $\mathcal{Y}_1$ and $\mathcal{Y}_2$ can be defined, according to [19], by $d_{\mathcal{G}}(\mathcal{Y}_1, \mathcal{Y}_2) = \sqrt{\sum_i \theta_i^2}$.

The metrics $d_{\mathcal{V}}$ and $d_{\mathcal{G}}$ are then used to compute the proposed *Geometric Motion History* features by analyzing sequentially the trajectories on the Stiefel and Grassmann manifolds, respectively. We note that data representation on these manifolds have been successfully used in literature and applied to human activity recognition [20], age estimation, and face recognition [21] from video sequences or sets of still images (for a more complete review, we refer the reader to [22]). Despite the common mathematical background, our methodology is quite different compared to the above-mentioned studies, as it maps depth videos to be trajectories on Grassmann manifold. This permits sequential analysis across the trajectories, making it possible to have a decision with lower latency (compared to full temporal observations [20]), and thus defining a stopping time.

## IV. EARLY EMOTION DETECTION USING SOSVM

The task of emotion detection is formulated as an early detection problem, which aims to detect the emotion of interest as quick as possible. This is achievd using SOSVM, which results into a convex optimization problem [23]. The main motivations for using SOSVM are: (1) it can be trained on all partial segments and the complete one at the same time; (2) it allows us to model the correlation between the extracted features and duration of the emotion; (3) no previous knowledge is required about the structure of the emotion; (4) it can give better performance than other algorithms in sequence-based applications [24].

### A. Extraction of Geometric Motion History Features

The underlying representations by trajectories on Stiefel or Grassmann manifolds allow us to use geometry tools to compute distances between points, thus quantifying the difference between successive low-resolution depth subsequences. Our idea here is to sequentially compute the distances between successive points and build a history of the body motion. More formally, given a trajectory $\mathcal{T}$ of $k$-dimensional subspaces of $\mathbb{R}^n$, $\{\mathcal{X}_i\}_{i \in \{0,\ldots,T\}}$, we compute sequentially the length of the geodesic connecting $\mathcal{X}_{i+1}$ to $\mathcal{X}_i$, which is added to the motion history of the fraction of video seen. This results into a one-dimensional signal varying along the time called *Geometric Motion History*, as depicted in Fig. 3. In particular, the plots show the *Geometric Motion History* feature vector obtained for three concatenated videos, where the green segment corresponds to the emotion of interest (*Happiness* in the Figure), which is comprised between two other emotions.

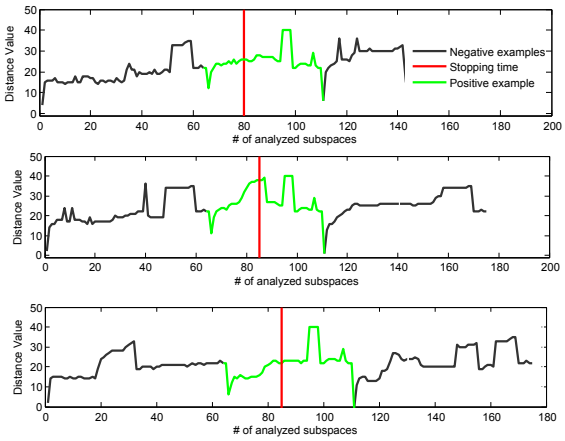


Fig. 3. Three examples of the *Geometric Motion History* feature vectors extracted using the proposed framework. It can be observed as the emotion of interest (*happiness*) extracted from different segments has a similar pattern, (highlighted in green) when randomly combined with different other emotion vectors. The red lines are the stopping times at which the early detection is performed online.

### B. Structured Output Learning from Sequential Data

Assume a set of concatenated *Geometric Motion History* feature vectors, $v_1, ..., v_n \in V$, as depicted in Fig. 3. Each resulted *Geometric Motion History* feature vector, $v_i$, includes only one emotion of interest, which is annotated by two values $s^i, e^i$, to define the start and the end of the emotion, respectively. At any instance within the emotion $s^i \leq t^i \leq e^i$, all partial emotions sub-segments obtained between $[1, t^i]$ will be used to train the SOSVM, since these different size sub-segments represent a positive state, whereas all other part of the *Geometric Motion History* vectors are negative. The expected performance from SOSVM in the testing stage is to fire the detection of the emotion of interest as soon as possible (after it starts and before it ends). We adopt similar methodology for sequential data learning by SOSVM to [13].

## V. EXPERIMENTS AND RESULTS

The proposed approach has been evaluated on the Cam3D Kinect database [25] using different scenarios and settings. In this database, Mahmoud et al. [25] collected a set of 108 audio/video segments of natural complex mental states of 7 subjects. Each video is acquired with the Kinect camera, including both the appearance (RGB) and depth (D) information. The data capture natural facial expressions and hand gestures accompanied. The emotional states are: *Agreeing*, *Bored*, *Disagreeing*, *Disgusted*, *Excite*, *Happy*, *Interested*, *Sad*, *Surprised*, *Thinking* and *Unsure*. These emotional states are more realistic and more complex than the basic well known six expressions in the literature. Table I shows the number of available segments for each emotional state.

It can be observed that videos in this dataset provide a sampling of the dimensional description chart of emotions as reported in Fig. 1. However, the possibility to use each emotion category in a detection experiment is hindered by the low number of videos comprised by several categories (i.e., less than 8 videos are present in 9 out of the 12 emotion categories, with 5 categories having just 1 or 2 videos). This motivated us to consider the following two experimental scenarios: *Happiness vs. others*; and *Thinking/Unsure vs. others*. Compared to the chart of Fig. 1, the first scenario tests the detection of an emotion located in the *high-arousal/pleasure* quadrant (positive emotion); the second one refers to an emotion in the *low-arousal/displeasure* sector (negative emotion).

TABLE I
NUMBER OF AVAILABLE VIDEOS FOR EACH EMOTIONAL STATE.

| Emotional/Mental State | # of segments |
|---|---|
| Agreeing | 4 |
| Bored | 3 |
| Disagreeing | 2 |
| Disgusted | 1 |
| Excited | 1 |
| **Happy** | **26** |
| Interested | 7 |
| Neutral | 2 |
| Sad | 1 |
| Surprised | 5 |
| **Thinking** | **22** |
| **Unsure** | **32** |

Three different evaluation criteria are used to test the performance from the viewpoint of accuracy and timeliness and the quality of emotion localization: (1) **Area under the ROC curve:** A ROC curve is created by plotting the True Positive Rate (TPR) vs. the False Positive Rate (FPR) at varying threshold; (2) **AMOC curve:** The Activity Monitoring Operating Characteristic curve is generally used to evaluate the timeliness of any event surveillance system; (3) **F1-score curve** or the F-measure which considers both the precision and the recall of the test to compute the score.

We applied the proposed framework to detect emotional states from two different regions of the emotion chart of Fig. 1: (1) *Happiness* out of all non-happiness (high-arousal/pleasure quadrant); (2) *Thinking/Unsure* vs. others (low-arousal/displeasure quadrant). The emotion of interest and other segments in the two experiments are divided equally into training and testing parts. Then, a concatenation of the computed *Geometric Motion History* features of each emotion of interest with other two signals from different emotional states in the training and testing sets is performed, as illustrated in Fig. 3. With this, we derive a total of 100 GMH for training, and the same number for testing. For each generated sequence, the onset and the offset point of the emotion of interest is known. The distances computed on Stiefel and Grassmann manifold (see Sec. III) are extracted for comparison. The effect of the window size used for extracting the sub-sequences of a segment is also analyzed.

For the *Happiness vs. non-happiness* case, Fig. 4 shows the ROC and the AMOC curves obtained using the *Geometric Motion History* feature computed for Grassmann and Stiefel manifold by averging the results of 20 different runs. From the ROC curves related to the Grassmann it can be seen that when the FPR is around 20% the TPR reaches 90% for *Happiness* detection. This accuracy decreases significantly

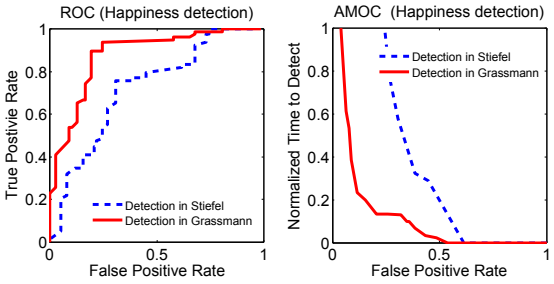


Fig. 4. ROC and AMOC curve for *Happiness* detection over Stiefel and Grassmann Manifolds.

(around 50%) at FAR=10%. Comparing the analysis of the trajectories along the Stiefel (dashed curves) and the Grassmann manifold (continued curves), it clearly emerges the sequential analysis performed on Grassmann manifold outperforms the analysis on Stiefel manifold. The areas under ROC curves are 0.73 and 0.84 on Stiefel and Grassmann, respectively. This demonstrates the consistency of the subspace based representation $\mathcal{Y} = Span(Y)$ and the associated metric $d_{\mathcal{G}}$ over the matrix representation. This is mainly due to the invariance of the subspace representation to the rotations $O(k)$ as $\mathcal{G}$ is a quotient space of $\mathcal{V}$ under the group action of $O(k)$. The plots on the right of Fig. 4 show the evolution of the system latency (the fraction of video needed to make the binary decision) against FPR. For example, the detector achieves 20% of FPR by analyzing 20% of the video segment. Once again, the results reported using the Grassmann representations are better compared to the Stiefel representation.

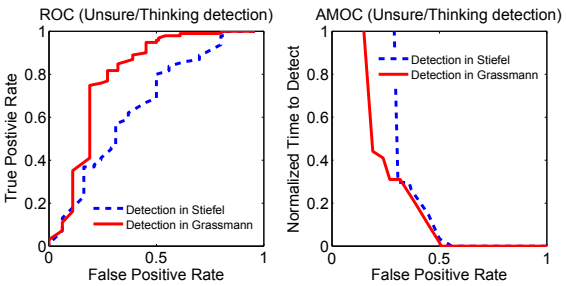


Fig. 5. ROC and AMOC curves for *Thinking/Unsure* detection over Stiefel and Grassmann manifolds.

In a second experiment, the detection accuracy for the *Thinking/Unsure* affective state is considered. Results reported in Fig. 5, show a performance decrease with respect to the Happiness detection (confirmed also by the F1-scores reported in Fig. 8). The areas under the ROC curves are 0.66 and 0.79 on Stiefel and Grassmann manifold, respectively. These results justify using the Grassmann rather Stiefel representation. From the plot on the right of this Figure, it can be noted that about 20% of the negative samples are recognized to be element of this class, even if the videos are completely seen. This can be motivated by the "common" neutral behavior exhibited by human beings when conveying other complex mental states (e.g., agreeing, bored, etc.). This induces a confusion to the detector, which was not the case for the previous Happiness detector, as the happiness is often accompanied by body and facial expressions.

To investigate the importance of using the upper part of the body versus using only the face, we conducted the same previous protocol on the database after cropping only the face region. From Fig. 6, it is clear that the upper part of the body expression is more informative than the facial expression in conveying the emotion of interest when filmed using cost-effective cameras. In *Happiness* experiment, the area under the ROC curve for the upper part of the body and the face only are 0.84 and 0.68, respectively. Following the same behavior, for the *Thinking/unsure* experiment, the area under ROC curve are 0.79 and 0.63 for the upper part of the body and the face only, respectively.

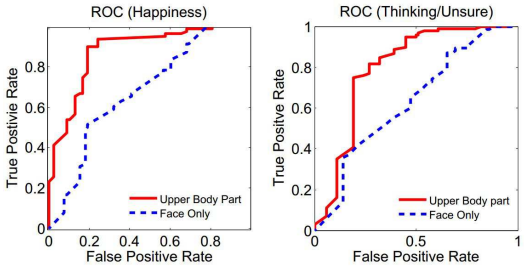


Fig. 6. ROC curves for *Happiness* and *Thinking/unsure* detection over the Grassmann manifold using the upper body and the face only.

To highlight the importance of the size of the window (# of frames used to embody the motion in the subspace) and the subspace dimension, in Fig. 7 we consider the Grassmann manifold for Happiness detection and compare results for windows of size $w$=20 and $w$=5 (red and blue curves, respectively). The dimension of the subspace is $k$=5 in both cases. In the first case, the window size of $w$=20 permits to keep 90% of the original information; in the second case ($w$=5), we keep 100% of the information as $k$=$w$=5. So, in this comparison the window size $w$ is the only changing parameter. The areas under the ROC curve for $w$=5 is 0.74, and 0.84 when $w$=20. The observed performance gap between the two cases (a quite marked improvement is noted for $w$=20), clearly evidences the importance of an appropriate setting of these parameters.

In a last experiment, we repeated 100-times the previous experiments with the optimal parameters ($w$=20 and $k$=5). In each run, the negative examples before and after the positive example are randomly selected, and the average F1-score (± standard deviation) is reported against the fraction of the video seen. Results are shown in Fig. 8, for the *Happiness* and *Thinking/Unsure* detectors (red and green curves, respectively). For short fractions of the event seen, the two cases show similar behavior, while the Happiness result clearly outperforms the Thinking/Unsure when the

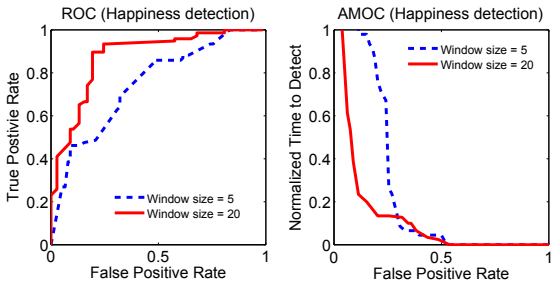


Fig. 7. ROC and AMOC curves for *Happiness* detection over the Grassmann manifold for two different window size (i.e., w=5 and w=20).

fraction of the event increases.

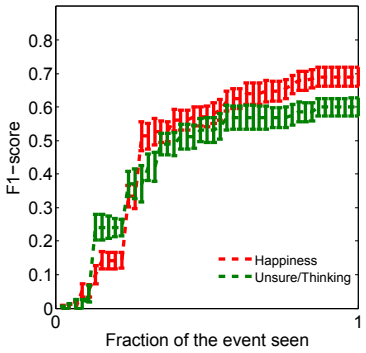


Fig. 8. Average F1-scores (with standard deviation) obtained for the *Happiness* emotion and the *Thinking/Unsure* affective states against the fraction of the event seen.

## VI. CONCLUSIONS

In this paper, We introduced new features (called *Geometric Motion History*) computed on human depth-videos for the purpose of early spontaneous emotion detection. Our idea is to consider well-established continuous emotion spaces (e.g., Valence-Arousal) and to define a region of interest for automatic detection. The dynamic depth-maps are first mapped to Grassmann manifold to face the quality of the data (low resolution, missing data, and noise). Then, the *Geometric Motion History* features are obtained by computing the velocity vectors along the trajectories on the Grassmann manifold. The use of methods dedicated to sequential analysis (such as SOSVM) allows us to study the trade-off between the detection accuracy and the system latency. Our approach has been tested on the new Cam3D dataset, which includes a limited number of annotated/segmented videos. Results evidence the viability of the approach for two specific emotional states (i.e., Happiness and Thinking/Unsure). It also demonstrates a clear benefit of using the expression of the upper part of the body, instead of face expression alone.

## VII. ACKNOWLEDGMENTS

We would like to thank the authors of [13] for providing the SOSVM codes and the evaluation tools. This work was supported in part by the MAGNUM 2 project (BPI and Région Nord-Pas de Calais).

## REFERENCES

[1] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.22, pp.1424–1445, 2000.

[2] Y. Sun, X. Chen, M. Rosato and L. Yin, "Tracking vertex flow and model adaptation for 3D spatio-temporal face analysis," *IEEE Trans. on Systems, Man, and Cybernetics – Part A*, vol.40, pp.461–474, 2010.

[3] B. Ben Amor, H. Drira, S. Berretti, M. Daoudi, and A. Srivastava, "4–D facial expression recognition by learning geometric deformations, *IEEE Trans. on Cybernetics*, vol.44, no.12, pp.2443,2457, Dec. 2014.

[4] P. Ekman, "Universals and cultural differences in facial expressions of emotion," in *Nebraska Symposium on Motivation*, Lincoln, vol.19, pp.207–283, 1972.

[5] J. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Jour. of Research in Personality*, vol.11, pp.273–294, 1977.

[6] A. Vinciarelli, M. Pantic and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing Jour.*, vol.27, pp.1743–1759, 2009.

[7] G. Sandbach, S. Zafeiriou, M. Pantic and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image and Vision Computing*, vol.30, pp.683–697, 2012.

[8] S. Wan and J. Aggarwal, "Spontaneous facial expression recognition: A robust metric learning approach," *Pattern Recognition Jour.*, vol.47, pp.1859–1868, 2014.

[9] M. Abd El Meguid and M. Levine, "Fully automated recognition of spontaneous facial expressions in videos using random forest classifiers," *IEEE Trans. on Affective Computing*, vol.5, pp.141–154, 2014.

[10] K. Schindler and L. Van Gool, "Action snippets: How many frames does human action recognition require?," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp.1–8, 2008.

[11] L. Su, S. Kumano, K. Otsuka, D. Mikami, J. Yamato and Y. Sato, "Early facial expression recognition with high-frame rate 3D sensing," in *IEEE Int. Conf. on Systems, Man, and Cybernetics*, pp.3304–3310, 2011.

[12] L. Su and Y. Sato, "Early facial expression recognition using early rankboost," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp.1–7, 2013.

[13] M. Hoai and F. De la Torre, "Max-margin early event detectors," *Int. Jour. of Computer Vision*, vol.107, pp.191–202, 2014.

[14] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Trans. on Affective Computing*, vol.4, pp.15–33, 2013.

[15] J. Van den Stock, R. Righart and B. de Gelder, "Body expressions influence recognition of emotions in the face and voice," *Emotion Jour.*, vol.7, pp.487–494, 2007.

[16] H. Meeren, C. van Heijnsbergen and B. de Gelder, "Rapid perceptual integration of facial expression and emotional body language," in *National Academy of Sciences*, USA, vol.45, pp.16518–16523, 2005.

[17] N. Bianchi-Berthouze, "Understanding the role of body movement in player engagement," in *Human Computer Interaction*, vol.28, pp.40–75, 2013.

[18] G. H. Golub and C. F. Van Loan, "Matrix Computations," (3rd Ed.), *Johns Hopkins University Press*, Baltimore, MD, USA, 1996.

[19] A. Edelman, T. A. Arias and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl*, vol.20, pp.303–353, 1998.

[20] P. K. Turaga, A. Veeraraghavan, A. Srivastava and R. Chellappa, "Statistical computations on grassmann and stiefel manifolds for image and video based recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.33, pp.2273–2286, 2011.

[21] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: A unifying view on subspace-based learning," in *Int. Conf. on Machine Learning*, vol.08, pp.376–383, 2008.

[22] Y. M. Lui, "Advances in matrix manifolds for computer vision," *Image Vision Computing Jour.*, vol.30, pp.380–388, 2012.

[23] I. Tsochantaridis, T. Joachims, T. Hofmann and Y. Altun, "Large margin methods for structured and interdependent output variables," *Jour. of Machine Learning Research*, vol.6, pp.1453–1484, 2005.

[24] N. Nguyen and Y. Guo, "Comparisons of sequence labeling algorithms and extensions," in *Int. Conf. on Machine Learning*, pp.681–688, 2007.

[25] M. Mahmoud, T. Baltrůsaitis, P. Robinson and L. Riek, "3D corpus of spontaneous complex mental states," in *Conf. on Affective Computing and Intelligent Interaction*, pp.205–214, 2011.