A Gaussian Mixture Representation of Gesture Kinematics for On-Line Sign Language Video Annotation

Fabio Martínez[1,2(✉)], Antoine Manzanera[2], Michèle Gouiffès[1], and Annelies Braffort[1]

[1] LIMSI, CNRS, Université Paris-Saclay, Paris, France
[2] U2IS/Robotics-Vision, ENSTA-ParisTech, Université Paris-Saclay, Paris, France
fabio.martinez-carillo@ensta-paristech.fr

Abstract. Sign languages (SLs) are visuo-gestural representations used by deaf communities. Recognition of SLs usually requires manual annotations, which are expert dependent, prone to errors and time consuming. This work introduces a method to support SL annotations based on a motion descriptor that characterizes dynamic gestures in videos. The proposed approach starts by computing local kinematic cues, represented as mixtures of Gaussians which together correspond to gestures with a semantic equivalence in the sign language corpora. At each frame, a spatial pyramid partition allows a fine-to-coarse sub-regional description of motion-cues distribution. Then for each sub-region, a histogram of motion-cues occurrence is built, forming a frame-gesture descriptor which can be used for on-line annotation. The proposed approach is evaluated using a bag-of-features framework, in which every frame-level histogram is mapped to an SVM. Experimental results show competitive results in terms of accuracy and time computation for a signing dataset.

1 Introduction

Sign languages (SLs) are natural languages used to communicate with and among the deaf communities, which, like spoken languages, differ from one country to another. Additionally, SLs are less-resourced languages with very few reference books describing them (grammar rules, etc.), a limited number of dictionaries and corpora, and even less dedicated processing tools.

Annotation software are tools used for linguistic studies, that allow researchers to visualize their data (mainly videos for SLs), annotate them with linguistic inputs, and analyze these inputs [1]. These corpus-based studies allow to create statistically-informed models that are useful for SL description, but also for SL processing. At this moment, such software are limited to only include automatic processing on the secondary data, the annotations, which are textual data. They do not include automatic processing on the primary data, the video. However, there is a growing interest on image and video processing tools, to characterize particular recorded gestures from local and global primitives such as motion, shape, body parts interactions, among others [2,3].

G. Bebis et al. (Eds.): ISVC 2015, Part II, LNCS 9475, pp. 293–303, 2015.
DOI: 10.1007/978-3-319-27863-6_27

This paper introduces a new proposal to support SLs annotations based on a motion descriptor that characterize temporal gestures in video sequences. The proposed approach is developed for French Sign Language corpus annotation. It starts by computing semi-dense trajectories, provided by point tracking in consecutive frames, over a set of gestures recorded in a video. Then, kinematic-cue words, represented as local mixture of Gaussians, are recursively computed at each time and for each trajectory during the video. These features are extremely fast to compute, and the action descriptor is available at each frame, thus allowing prediction on partial video sequences, and then on-line gesture recognition capability.

2 Sign Language

2.1 Main Linguistic Properties

SLs are visuo-gestural representations that follow specific rules induced by use and interaction among corporal articulators and the visual perception. This language promotes the simultaneous use of a number of articulators, the linguistic use of the space in front of the signer so-called 'signing space', and the omnipresence of iconicity at all levels of the language [4]. The main linguistic specificities and challenges are the followings:

- Signs can be broken down into smaller constituents whose linguistic nature, definition and detection are still subject to debate;
- Signs can bear strong modification of their constituents depending on the context, and modelling all possible variations can require too many different training examples to keep the categories consistent;
- Signs can be more or less lexicalised, and the most productive ones are built on the fly and are not indexed in a dictionary, which makes them extremely difficult to be modelled from a classical approach.

SLs characterization must also consider non-manual activity that convey meaningful information. For instance, SL production involves non-manual articulators such as head, face, and torso which are relatively synchronized on different spatial and temporal scales. In fact, the signer uses *the signing space* to support and topologically structure his discourse. This spatial and multi-component property, as well as the importance of the productive signs make the design of SL processing tools a very challenging task.

2.2 The LSF (French Sign Language) Corpora

The corpora used in this study is extracted from the corpus collected during DictaSign, a three-year FP7 ICT project that aimed to improve the state of web-based communication for deaf people [5]. It is composed of nine videos that contain isolated dates, such as *'Lundi 2 novembre 2013'* (Monday, November 2nd, 2013). The lexicon is constituted of the seven days, the twelve months, and

a set of numbers. In LSF, a date is composed of four elements following the order: DAY NUMBER MONTH YEAR. The day and month signs are simple gestural units than can present regional variants. These dates are less complex than SL utterance, but they include various issues such as lexicon variability, co-articulation. They also include some spatial constraints, but limited to the image plane. This seems to us good candidates as a first step, to evaluate the performance of our method on motion description with this kind of data.

3 Background on Motion Descriptors

Motion analysis is a fundamental tool to segment potential region of interest, quantify, detect, recognize gestures or describe spatio-temporal interactions. One of the advantages of motion based characterization is the relative independence to appearance, which has potential applications in uncontrolled conditions.

Motion descriptors based on tracked local space-time trajectories from optical flow fields, currently provide the best performance to represent gestures and understand video sequences [6–8]. To recognize human activities, these strategies namely integrates local features along the trajectories to capture shape, appearance and motion information, namely using HOF (Histograms of Optical Flow), MBH (Motion Boundary Histograms) and HOG (Histograms of Oriented Gradients) [7]. This descriptor was also used in [9], but using improved motion trajectories obtained by correcting the camera motion in video sequences. In general, these descriptors are dependent on the appearance and structural image features computed around trajectories, fact that could be critical in language recognition in which shape signs and appearance have high inter subject variability. Additionally, the spatio-temporal volumes are heuristically cut off from a fixed temporal length (for example: 15 frames in [7]) that may be a problem to represent series of gestures of a SL utterance that can vary from one subject to another and also depending of the represented dialog. Besides, the dynamic trajectory information is poorly exploited, *i.e.*, the action descriptors only use the trajectory as information support to compute static frame features (namely spatial features such as image gradients), neglecting relevant kinematic information that is naturally available on the trajectory.

Other works have characterized the dynamic of dense beams of trajectories to describe actions in video sequences. For instance, in [10] a set of k cut trajectories are characterized using first order derivatives in the (x, y) axes, which may be sensitive to motion direction and to scale. In [11] is firstly considered strong sparse coding assumptions to filter out motion trajectories. Then, the remaining trajectories are characterized using *Largest Lyapunov Exponent* and the *correlation dimension*.

Specifically, in the domain of SLs, several works have been focused in the automatic recognition of atomic gestures by characterizing postures, shape regions, global movements among others (see [3] for an overview of the domain). These works include the use of a broad spectrum of methods such as tracking of articulated shapes, colour segmentation to characterize postures, and the

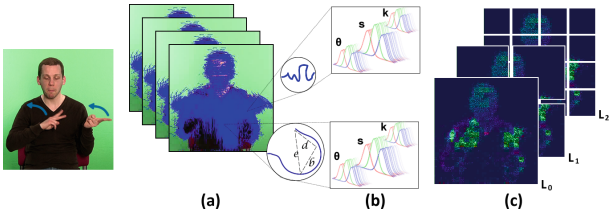

Fig. 1. Pipeline of the proposed approach for SL recognition and annotation support. (a) First, a set of trajectories are computed. (b) For each trajectory, a set of kinematic-cue words are computed using recursive Mixture of Gaussian. (c) A pyramidal partition is applied at each frame to support hierarchical BoW representation, which is in turn used to recognize particular FSL gestures.

static and temporal characterization shape articulators. In terms of annotation support, some projects have tried to integrate video analysis modules into annotation software. For instance, *Ancolin* is a prototype annotation software [12] developed onto a distributed architecture, that includes several external plugins for sign language video processing such as colour skin detection, characterization of head shape and size, and motion history images to code arm movements. This video characterization provides additional useful information to the annotation but remains dependent on accurate segmentation of human silhouettes and is also highly dependent on the user. *SignStream* [1] is another annotation software, currently used for linguistic analysis that includes components for 3D head detection and tracking to estimate head gesture: Currently, this application includes new modules to automatically characterize hand gestures in ASL using a tracking system [2]. Additionally, the *SLMotion* toolkit provides a framework for automatic and semi-automatic analysis, feature extraction and annotation of individual sign language videos. The program includes support for exporting the annotations in ELAN[1] format.

4 The Proposed Method

The proposed strategy recognizes SL gestures by using an on-line spatio-temporal characterization of the signer movements recorded in a video. The Fig. 1 illustrates a pipeline of the proposed approach.

4.1 Computing Semi-dense Trajectories

Point trajectories are useful motion features based on tracking salient points along the video sequence, allowing in most cases a relevant representation of action present in the video. The proposed approach requires a set of trajectories with a suitable trade-off between accuracy and computation time, in order

[1] Software widely used for linguistic analysis of video data.

Fig. 2. Spatio-temporal representation of trajectories and MoG kinematic representation. Each row represents a spatio-temporal gesture corresponding to two days. The second and third columns illustrate the dense trajectories and their kinematic descriptors, respectively. The fourth and fifth columns correspond to the semi-dense trajectories and their kinematic descriptors.

to support fast annotation prediction. In this work were considered two different methods to compute motion trajectories (see examples in Fig. 2), described hereunder:

Dense trajectories [7] are extracted from a dense optical flow field estimated at multiple spatial scales and regularized using a median filter. Additionally, a trajectory is considered as outlier and removed if it meets any of the two following conditions: (1) the standard deviation of the velocity along the trajectory is above a given threshold, and (2) it presents sudden displacements, corresponding to vectors whose magnitude is larger than a certain proportion of the overall displacement of the trajectory.

Semi-Dense trajectories [13] are computed from a set of weakly salient points, tracked using a coarse-to-fine prediction and matching approach, allowing a high degree of parallelism and dominant movement estimation. This technique produces high density trajectory beams, robust to large camera accelerations and allowing statistically significant trajectory based representation, with a good trade-off between accuracy and performance.

4.2 Gaussian Mixture Representation of Kinematic Features

Each computed trajectory $\boldsymbol{\Gamma}(t) \in \mathbb{R}^2$ represents a particle traveling in the 2d space (x, y) from time t_1 to t_n. At each time t, the trajectory motion information can be characterized by a collection of kinematic features $\{F_t^i\}_i$, such as the velocity, acceleration, curvature among others, using finite difference approximation. In this work, each computed kinematic feature is modelled as a random variable following a mixture of K Gaussian densities, whose parameters are defined as: $\sum_{k=1}^{K} w_t^k \mathcal{N}(\mu_t^k, \sigma_t^k)$, where (μ_t^k, σ_t^k) are the mean and standard deviation of each Gaussian mode and w_t^k represents the contribution of each mode, with $\sum_{k=1}^{K} w_t^k = 1$.

Algorithm 1. Recursive Mixture of Gaussian estimated locally for each computed trajectory and for each considered kinematic feature.

for each time t **do**
  for each trajectory Γ **do**
    if $|\Gamma| > 3$ **then**
      for each kinematic feature i **do**
        calculate feature $F_t = F(i)$
        for each mode k **do**
          if $\left|F_t - \mu_t^k\right| \leq \lambda\sigma_t^k$ **then**
            $\mu_t^k = \mu_{t-1}^k + \alpha(F_t - \mu_{t-1}^k)$
            $(\sigma_t^k)^2 = \alpha\left(\mu_{t-1}^k - F_t\right)^2 + (1-\alpha)\,(\sigma_{t-1}^k)^2$
            $\omega_t^k = \omega_{t-1}^k + \alpha$
          end if
        end for
        normalize ω_t^k **such that** $\sum_{k=1}^{K} w_t^k = 1$
        rank the modes in decreasing order of $\frac{\omega_t^k}{\sigma_t^k}$
        keep the first B **modes such that:** $B = \arg\min_{b=1}^{K}\left\{b; \sum_{k=1}^{b} w_t^k > \mathbf{T}\right\}$
      end for
    end if
  end for
end for

The MoG representation is herein implemented as described in Algorithm 1 [14]. This algorithm allows an on-line MoG updating and therefore a kinematic gesture representation is available at each frame. First, the density parameters are initialized, assigning to the mean the first value of each kinematic feature computed, to the standard deviation any fixed value and the weight $\omega_{t,k}$ being the same for each mode k. Then, the distributions that are most likely matched by the current kinematic sample (i.e. when the sample distance to the mode is less than λ times its standard deviations) are updated. The density parameters $\left\{\mu_t^k, \sigma_t^k\right\}$ are updated using a on-line cumulative filter with a learning rate parameter $\alpha \in [0, 1]$ which takes into account the history of the kinematic measure along the trajectory, with $t \approx 1/\alpha$. Each ω_t^k is also updated according to the matched distribution at each time. After that, the distributions are sorted in decreasing ordered according to $\frac{\omega_t^k}{\sigma_t^k}$. Finally, only the B first distributions of the MoG are considered. If any distribution is initialized (no existing one is matched), then the parameters of the distribution with lowest weight are replaced by the initial values. This recursive representation has the main advantage of computational speed which is essential to on-line annotation tools, the recent history of each kinematic measure being available at each frame.

Kinematic Features $\boldsymbol{F}_t$: In order to keep the computation fast, the kinematic features considered in this work were: the velocity $\boldsymbol{v}(t) = \boldsymbol{\Gamma}'(t)$, depicted by its direction $\theta(t) = \arg \boldsymbol{v}(t)$ and modulus (speed) $s(t) = ||\boldsymbol{v}(t)||$. The curvature was also included; it is related with how rapidly the trajectory is bending to one side, and corresponds to the normal acceleration when the curvilinear speed is constant. The curvature is herein implemented as proposed in [15], using finite difference on consecutive points of Γ as follows: $\kappa(t-1) = \frac{\sqrt{\zeta(\zeta-b)(\zeta-d)(\zeta-e)}}{bde}$, where $\zeta = (b+d+e)/2$, as illustrated in Fig. 1.

Each trajectory is then characterized at time t by the set of kinematic features $\boldsymbol{F}_t = \{\theta(t), s(t), \kappa(t)\}$. The proposed strategy is flexible to include any other local kinematic measure computed along a 2d trajectory. An additional advantage of the proposed strategy is that any kinematic feature can access independently to the recognition or a set of features can be chosen through a learning stage, in order to reach higher execution times or reduce memory requirements, preserving a proper accuracy. Figure 2 shows computed trajectories. The recursive means of computed kinematic features are represented using a RGB color map representation, the blue being the curvature and the red and green being respectively the modulus and the direction of the velocity.

4.3 Spatial Pyramid Representation and Codebooks Learning

In SLs, the signs are visuo-gestural representations, which require a temporal and spatial characterization. In the proposed approach, a regional analysis of the MoG features are carried out by following a fine-to-coarse partition of each frame. This *spatial pyramid* forms a set of partition layers $\{L_i\}_{0 \leq i \leq N_r}$ (see in Fig. 1-(c)), whose total number of sub-regions is: $s\_r = \sum_{i=0}^{N_r} 4^i = \frac{4^{N_r+1}-1}{3}$.

In the training step, different configurations of the spatial pyramid representation were used to learn the codebooks of kinematic words computed from the MoG recursive representation. Each codebook is made up by a set of MoG kinematic words, formed by the output of a classical k-means algorithm computed using a random selection of 10 % of the MoG features extracted over the whole training video set. All the feature words are computed with the same α, with n the number of kinematic measures estimated at each time on each trajectory and b the number of modes retained from the MoG distribution. Then, each codebook contains k_l representative feature words, each word having a dimension of $3nb$. During a first configuration, a global codebook $\{D_0\}$ was learned from the region of L_0, and then a histogram of motion word occurrences was considered for each sub-region of the spatial pyramid. This histogram is constructed by counting the number of times each one of the k_l kinematic centroid is closest to the computed features, based on the Euclidean distance on $\mathbb{R}^{3n}$. In this case, the total size of the descriptor is the concatenation of histograms computed for each sub-region, with size of $s\_r \times k_{l_0}$. In a second configuration, for each sub-region of the spatial pyramid representation was considered a independent codebook. From the set of codebooks $\{D_l\}_{1 \leq l \leq \Lambda}$ is then computed histogram of occurrences with variable size according to the size of each regional codebook, resulting a more

compact descriptor w.r.t the first version. Finally, the labeling of each potential sign gesture is performed by a Support Vector Machine (SVM) using the standard LIBSVM [16] implementation, using the *one-against-one* multi-class SVM classification with a Radial Basis Function (RBF) kernel.

5 Results

A first exploration over a SL corpus of signatures representing *Dates* was carried out to evaluate the proposed approach in the task of sign recognition to support annotation. The experimental evaluation was performed under a leave-one-out cross validation scheme by using different segments of the videos. The best performance of the proposed approach was obtained with a pyramid of $N_r = 2$ levels and a learning rate of $\alpha = 0.25$ corresponding, to a time depth of 4 frames. The number of estimated modes in each MoG was set to 7, taking into account the 5 dominant modes.

Evaluations over the SL dataset were carried out taking into account different lexical complexities of the signs. First, the *words* recognition related with days and months was performed. Because the approach is based on statistical representations of spatio-temporal gestures, it was only considered gestures with more that 5 samples available into the dataset, corresponding to 4 days and 5 months. In Table 1 is shown the performance obtained by the proposed approach for recognizing these spatio-temporal gestures. In general the proposed approach is able to recognize different atomic gestures that correspond to localized movements. The best performance of the proposed approach was achieved by using dense trajectories with a compact spatial pyramid representation of regional dictionaries. Some mistakes in the recognition may be attributed to regional variations of gestural signs.

Second, the performance of the approach to recognize *dates* was evaluated. Complete dates have more complex lexical structure and they are composed

Table 1. Classification rate of individual gestures corresponding to days and months

Gesture	Spatial Configuration	Trajectories	
		Dense	Semi-dense
Days	Pyr single Dic (L_0)	74.07	66.66
	Pyr mult Dics	81.48	74
	without Sub-regions	70.37	62.96
–			
Months	Pyr single Dic (L_0)	75.05	63.45
	Pyr mult Dics	80	72.21
	without Sub-regions	65	55.3

Table 2. Classification rate using different spatial configurations and trajectories for complete date phrases

Spatial Configuration	Trajectories	
	Dense	Semi-dense
Pyr single Dic (L_0)	75.03	67.13
Pyr mult Dics	77.21	70.3
Without Sub-regions	72.45	62

by the ordered sign information of day, month and year[2]. The proposed approach achieves a recognition rate of 75 % on a total of 7 different dates. In Table 2 is shown the results obtained by using different spatial configurations and the different types of trajectory. The best results is obtained using a spatial pyramid configuration of multiple dictionaries learned by region and the dense trajectories. The both pyramidal representations herein implemented allows a more robust representation than a global space-frame description, i.e., without a sub-regional division. Some mistakes are due to the natural variability between different signers and to the limited number of samples available for each date.

Action Recognition Evaluation: Because the proposed approach is based on the recognition of spatio-temporal patterns, it can be extended to recognize other motion activities. An additional evaluation was herein considered in public action recognition datasets. Two different datasets were considered: the *KTH* (six action classes, contained in a total of 2 391 videos) and the *UT-Interaction* (six different interactions in 120 videos) [17]. Table 3 summarizes the results obtained for the proposed approach with other state-of-the art approaches. It generally achieves competitive results with the great advantage

Table 3. The right table reports the comparison with state-of-the-art methods using the KTH database following the original experimental setup [18]. The figures marked with (*) have been computed using a k-fold validation with $k = 5$ [19]. The left table shows the comparison with state-of-the-art methods using the UT-database using k-fold validation with $k = 10$, as described in [17]

Methods	Accuracy
Proposed approach	**92.23**
Wang *et. al.* [7]	94.2
Laptev *et. al.* [20]	91.8
Proposed approach *	**97.0**
Liu *et. al.* *[19]	93.8

Methods	Accuracy
Proposed approach	**90.3**
Laptev *et. al.* [6]	87.6
Yu *et. al.* [21]	83.3
Daysy [22]	71

[2] An example of a considered date is: *Vendredi douze septembre mille six cent quatre vingt dix*, which means Friday, September the 12^{th}, 1690.

of being computationally efficient and usable in real-time applications. In contrast, other motion descriptors typically use a lot of features for each trajectory, including appearance information.

The proposed approach achieved memory efficiency, taking in average 0.30 milliseconds for each frame to build the descriptor. The experiments were carried out on a single core i3-3240 CPU @3.40GHz.

6 Conclusions

This work introduced a new motion descriptor that is able to recognize motion gestures related with SLs. The proposed approach allows an on-line support to SL annotation, by combining trajectory beams and Mixture of Gaussian representation of kinematic cues. The motion cues are spatially aggregated at each frame using a pyramid representation. The proposed approach can also be included as a plugin in SL systems and used as part of more sophisticated SL analysis. A more exhaustive evaluation with a larger dataset will be performed in order to increase the statistical samples of each gesture.

Acknowledgements. This research is funded by the RTRA Digiteo project MAPOCA.

References

1. Neidle, C.: Signstream: A database tool for research on visual-gestural language. Sign Lang. Linguist. **4**, 203–214 (2001)
2. Gavrilov, Z., Sclaroff, S., Neidle, C., Dickinson, S.: Detecting reduplication in videos of american sign language. In: 5th Workshop on the Representation and Processing of Sign Language (RPSL) (2014)
3. Cooper, H., Holt, B., Bowden, R.: Sign language recognition (2011)
4. Braffort, A., Filhol, M.: Constraint-based sign language processing. Constraints and Language. Cambridge Scholar Publishing, Cambridge (2014)
5. Matthes, et. al: Elicitation tasks and materials designed for dicta sign's multilingual corpus. In: 4th Workshop on the Representation and Processing of Sign Language (RPSL 2010) of (LREC 2010)
6. Kantorov, V., Laptev, I.: Efficient feature extraction, encoding and classification for action recognition. cvpr (2014)
7. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. CVPR 2011, Washington, DC, USA, pp. 3169–3176. IEEE Computer Society (2011)
8. Jain, M., Jegou, H., Bouthemy, P.: Better exploiting motion for better action recognition. In: CVPR. CVPR 2013, pp. 2555–2562 (2013)
9. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV 2013, Sydney, Australia, pp. 3551–3558. IEEE (2013)
10. Matikainen, P., Hebert, M., Sukthankar, R.: Trajectons: Action recognition through the motion analysis of tracked features. In: ICCV Workshops, pp. 514–521. IEEE (2009)

11. Wu, S., Oreifej, O., Shah, M.: Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In: ICCV 2011, pp. 1419–1426 (2011)
12. Braffort, A., Choisier, A., Collet, C., Dalle, P., Gianni, F., Lenseigne, B., Segouat, J.: Toward an annotation software for video of sign language, including image processing tools and signing space modelling. In: LREC 2004 (2004)
13. Garrigues, M., Manzanera, A.: Real time semi-dense point tracking. In: Campilho, A., Kamel, M. (eds.) ICIAR 2012, Part I. LNCS, vol. 7324, pp. 245–252. Springer, Heidelberg (2012)
14. Gaber, M.M., Stahl, F., Gomes, J.B.: Background. In: Gaber, M.M., Stahl, F., Gomes, J.B. (eds.) Pocket Data Mining. SBD, vol. 2, pp. 7–22. Springer, Heidelberg (2014)
15. Boutin, M.: Numerically invariant signature curves. Int. J. Comput. Vis. **40**, 235–248 (2000)
16. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27:1–27:27 (2011)
17. Ryoo, M.S., Aggarwal, J.K.: UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA) (2010)
18. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: ICPR 2004. ICPR 2004, Washington, DC, USA, pp. 32–36 (2004)
19. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the wild". IEEE ICVPR (2009)
20. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Conference on Computer Vision and Pattern Recognition (2008)
21. Yu, T.H., Kim, T.K., Cipolla, R.: Real-time action recognition by spatio-temporal semantic and structural forest. BMVA Press **52**(1-52), 12 (2010)
22. Cao, X., Zhang, H., Deng, C., Liu, Q., Liu, H.: Action recognition using 3d daisy descriptor. Mach. Vision Appl. **25**, 159–171 (2014)