Motion Segmentation using Visual and Bio-mechanical Features

Stefano Alletto, Giuseppe Serra, Rita Cucchiara
University of Modena and Reggio Emilia, Department of Engineering
{*name.surname*}*@unimore.it*

ABSTRACT

Nowadays, egocentric wearable devices are continuously increasing their widespread among both the academic community and the general public. For this reason, methods capable of automatically segment the video based on the recorder motion patterns are gaining attention. These devices present the unique opportunity of both high quality video recordings and multimodal sensors readings. Significant efforts have been made in either analyzing the video stream recorded by these devices or the bio-mechanical sensor information. So far, the integration between these two realities has not been fully addressed, and the real capabilities of these devices are not yet exploited. In this paper, we present a solution to segment a video sequence into motion activities by introducing a novel data fusion technique based on the covariance of visual and bio-mechanical features. The experimental results are promising and show that the proposed integration strategy outperforms the results achieved focusing solely on a single source.

General Terms

Algorithms, Design, Experimentation

Keywords

Egocentric vision, wearable devices, multimodal sensor analysis

1. INTRODUCTION

Due to the wide spreading of head-mounted cameras, systems dealing with an egocentric perspective are arousing a growing interest in industries and in the research community. Egocentric videos captured with wearable cameras are long and unstructured, and their continuous nature yields no evident shot boundaries. Information such as events and personal experiences cannot be comfortably reviewed but require a manual search through the video sequences, which can be several hours long. For this reason, the need for automated tools that assist users in accessing to the information in such videos is clear.

Several works tackle with problems such as activity and gesture recognition, social interactions or object recognition exploiting the unique perspective of wearable cameras [6, 1, 2]. Although these approaches are robust enough in specialized contexts and short activities they are not designed to analyze long and unconstrained videos in order to identify interesting segments of video that contain relevant information.

Recently, Lu and Grauman [10] handled egocentric video summarization partitioning videos into relevant sub-shots on the basis of motion features analysis. They segment the video in three classes: *static*, *moving the head*, *in tansit* and smooth the classification results with a Markov Random Field. Kitani *et al.* [8] present an unsupervised learning approach that uses motion-based histograms in order to classify ego-action categories. Poleg *et al.* [12] propose to temporally segment an egocentric video into twelve long time activities (such as *walking*, *running* and *wheels*) using classifiers trained on feature vectors derived from the cumulative displacement curves.

While these methods address a key problem of egocentric video, they are doing so disregarding a major feature of many modern wearable devices. From smartphones to Google Glass, many of these devices equip bio-mechanical sensors such as accelerometers and gyroscopes. In fact, the low costs, low power requirements and small dimensions of these sensors make them perfect for embedded and wearable applications. A new range of applications and techniques arise from their use: Li *et al.* [9] employ gyroscope and accelerometer to detect the posture of a subject, focusing on harmful situations like falls. Recently, the high precision of these sensors has been exploited by Hernandez *et al.* [7] to predict in real-time physiological parameters of their user such as heart and respiration rates. Differently, Brunetto *et al.* [3] exploit sensor information in a SLAM framework where accelerometer and gyroscope are combined with a RGB-D camera and are employed to enforce the orientation estimation, effectively improving the keyframe extraction process.

In this paper, we propose a novel approach to integrate a robust visual motion descriptor, namely Optical Flow (the key descriptor exploited in the aforementioned ego-vision video segmentation approaches [10, 8, 12]), with heterogeneous sensor data. In particular, our proposal is to build a descriptor based on the covariance of visual and iner-

MM '16 Amsterdam, The Netherlands

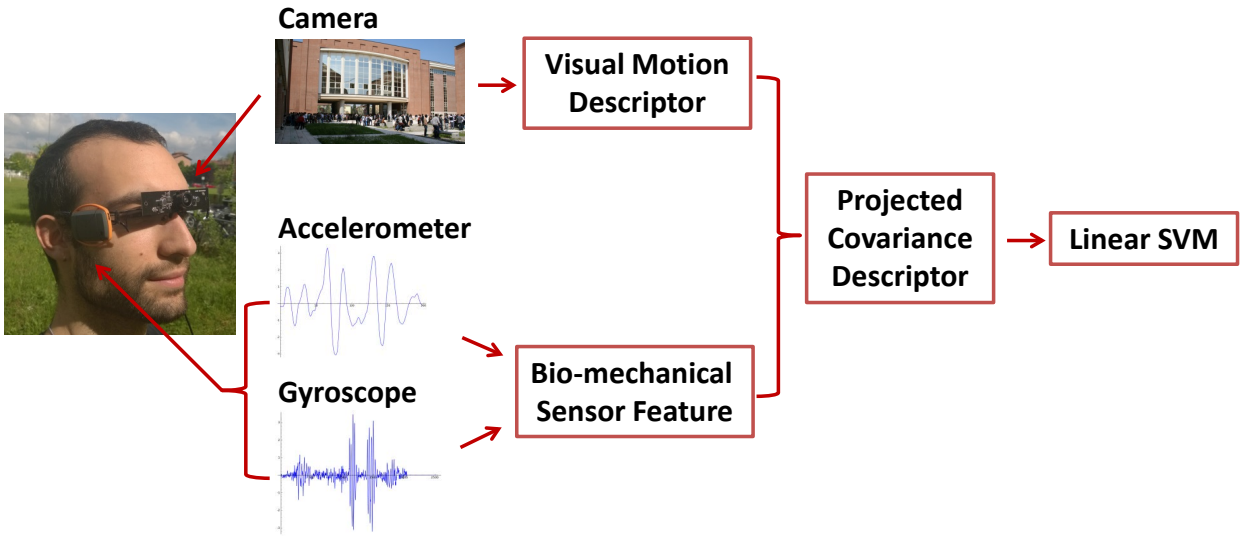


Figure 1: A schematization of the proposed solution.

tial information. In contrast with common early fusion approaches which directly combine (e.g concatenating the resulting feature vectors) heterogeneous data, our strategy allows to capture the correlation between visual and inertial features. The results obtained are encouraging and demonstrate clear benefits in the integration between visual and bio-mechanical features.

Furthermore, due to the novelty of the proposed study, we contribute to further research efforts on the matter releasing both the dataset and the source code. The dataset, referred as EGO-SENSORS in the following, features several hours of egocentric video coupled with tri-axial accelerometer and gyroscope data, fully annotated at a frame level with the current motion label.

2. MOTION SEGMENTATION

To segment egocentric videos into motion classes, we analyze the visual and bio-mechanical data stream coming from a glass-mounted wearable device, which features a camera, a tri-axial accelerometer and a tri-axial gyroscope. The Figure 1 depicts a schematization of our solution.

In particular, we analyze the multimodal stream using a sliding window approach. Two different sets of windows are considered: a smaller window $\mathbf{W} = \{\mathbf{f}_1, \ldots, \mathbf{f}_N\}$ that aggregates motion features for each frame, and a larger one $\mathbf{S} = \{\mathbf{W}_1, \ldots, \mathbf{W}_M\}$ that aggregates windows (N and M are the size of the two windows respectively). The frame-level window W allows to compute metrics over small temporal intervals: this choice balances the trade-off between fine information grain and noise, and the 50% overlap between windows has demonstrated to improve sampling robustness [13]. For each window W, motion features are extracted from the visual content and from the tri-axial accelerometer and gyroscope respectively (see subsection 2.1 and subsection 2.2). The windows $\mathbf{S}$ aggregates small windows into larger and more temporally relevant sequences spanning a few seconds of video.

To deal with these heterogeneous data we propose an approach that efficiently integrates and captures the correlation between them. In particular, let $\mathbf{S}$ be a set of descriptors extracted on a large window, where $\mathbf{W}_i = [\mathbf{mv}_i \mathbf{mb_i}]$ is the concatenation of heterogeneous features of the sub-window $\mathbf{W}_i$, we represent them by a covariance matrix $\mathbf{C}$. It encodes information about the variance of the features and their correlations and is computed as follows:

$$\mathbf{C} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{W}_i - \mathbf{m})(\mathbf{W}_i - \mathbf{m})^T, \tag{1}$$

where $\mathbf{m}$ is the mean vector of the set $\mathbf{W}_i$. Although the space of covariance matrices can be formulated as a differentiable manifold, it does not lie in a vector space (e.g the covariance space is not closed under multiplication with a negative scalar) and Euclidean distance between image descriptors can not be computed. Therefore to use this descriptive feature vector, we need to define a suitable transformation. We exploit a projection from the Riemannian manifold to an Euclidean tangent space, called Log-Euclidean metric as suggested by [14]. The basic idea of the Log-Euclidean metric is to construct an equivalent relationship between the Riemannian manifold and the vector space of the symmetric matrix.

The first step is the projection of the covariance matrix on an Euclidean space tangent to the Riemannian manifold, on a specific tangency matrix $\mathbf{T}$. The second one is the extraction of the orthonormal coordinates of the projected vector. In the following, matrices (points in the Riemannian manifold) will be denoted by bold uppercase letters, while vectors (points in the Euclidean space) by bold lowercase ones. The projection of $\mathbf{C}$ on the hyperplane tangent to $\mathbf{T}$ becomes:

$$\mathbf{c} = \mathrm{vec}_{\mathbf{I}} \left(\log \left(\mathbf{T}^{-\frac{1}{2}} \mathbf{C} \mathbf{T}^{-\frac{1}{2}} \right) \right), \tag{2}$$

where log is the matrix logarithm operator and $\mathbf{I}$ is the identity matrix, while the vector operator on the tanget space at identity of a symmetric matrix $\mathbf{Y}$ is defined as:

$$\text{vec}_{\mathbf{I}}(\mathbf{Y}) = \left[y_{1,1}\ \sqrt{2}y_{1,2}\ \sqrt{2}y_{1,3} \dots y_{2,2}\ \sqrt{2}y_{2,3} \dots y_{d,d} \right]. \tag{3}$$

As observed in [11], by computing the sectional curvature of the Riemmanian manifold, the natural generalization of the classical Gaussian curvature for surfaces, it is possible to show that this space is almost flat. This means that the neighborhood relation between the points on the manifold remains unchanged, wherever the projection point $\mathbf{T}$ is located. Therefore, from a computational point of view, the best choice for $\mathbf{T}$ is the identity matrix, which simply translates the mapping into applying the $\text{vec}_{\mathbf{I}}$ operator to the standard matrix logarithm. This also frees us from the problem of optimizing the projection point for the specific data under consideration, leading to a generally applicable descriptor. Since the projected covariance is a symmetric matrix of $d \times d$ values, the image descriptor is a $(d^2 + d)/2$-dimensional feature vector.

The feature vectors resulting from this analysis are extremely descriptive and can be used to train a linear SVM classifier. This is particularly favorable due to the considered setting of embedded wearable devices that often features limited hardware capabilities and real-time processing needs.

2.1 Visual Motion Descriptor

To represent the visual components of apparent motion in each sub-window $\mathbf{W}$, we extract features based on optical flow and blurriness following [10]. The optical flow descriptor can both measure forward travel and effectively capture the fast head movements that so often occur in ego-vision. We compute dense optical flow for each couple of consecutive frames, using Farneback algorithm [5], obtaining the relative apparent velocity of each pixel (V_x, V_y). These values can be expressed in polar coordinates (magnitudes and orientations) as in the following:

$$M = \sqrt{V_x^2 + V_y^2} \quad \theta = \arctan(V_y / V_x) \tag{4}$$

We quantize both the orientations and magnitudes into eight bins. The final visual feature vector is then obtained concatenating the magnitude and the orientation histograms, where the latter is weighed by its magnitude.

To compute the blurriness descriptor we adopt the solution proposed by Roffet *et al.* [4], which evaluating the line and row difference between the original image and the image obtained applying to it a horizontal and a vertical strong low-pass filter.

The visual motion feature vector, i.e. the concatenation of the two visual descriptors, is used to compute the frame window W descriptor applying an *average pooling* strategy followed by a l^2 normalization.

2.2 Bio-mechanical Sensor Feature

The bio-mechanical descriptor is composed by data obtained from the accelerometer and the gyroscope. The raw data consists of three values corresponding to the head acceleration along x-axis, y-axis and z-axis expressed in $\frac{m}{s^2}$ and other three values that represent the orientations (azimuth, pitch and roll) expressed in $\frac{degree}{sec}$. However, classifiers are demonstrated to perform poorly on raw sensor data and require a representation that captures the predominant characteristics that are intrinsic to this kind of data [9]. In fact, Figure 2 shows a plot of the raw data obtained from the x-axis of the accelerometer. It can be seen how, aside from two peaks, no clear and discriminative information can be directly inferred from such data. Therefore, for each window W we compute different features for both data sensors: minimum and maximum values inside the window; mean and standard deviation of acceleration and orientation values along the three axes; amount of zero-crossings inside the window, that is the amount of changes in sign of the considered metric. The dimensionality of this descriptor can vary depending on which of these features are considered and is $30 - dimensional$ when all of them are employed together.

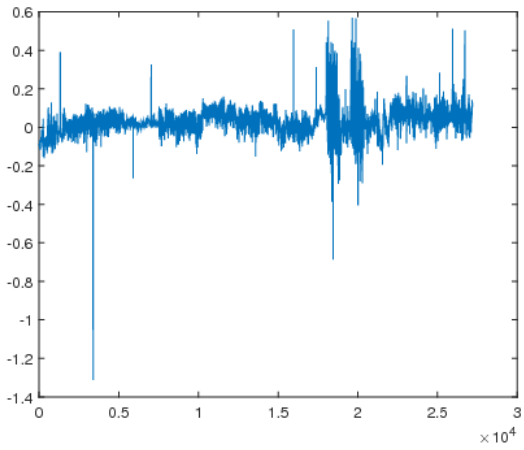


Figure 2: An example of raw data acquired from the x-axis of an accelerometer over 25.000 frames.

3. EXPERIMENTAL RESULTS

To thoroughly evaluate the proposed approach, we record the novel dataset EGO-SENSORS. It was collected by using glass-mounted device that consists of a tri-axial accelerometer and a tri-axial gyroscope embedded in the EXLs1 sensor node, a camera and an Odroid-XU developer board used as data-processing and storage unit.

The developer board we use embeds the ARM Exynos 5 SoC, that hosts a Quad big.LITTLE ARM processor (Cortex A15 and A7). Since the wearable camera acquires data at 30 frames per second and the EXLs1 sensor streams at 100 Hz, we synchronize the two by sub-sampling the sensor readings at 30 Hz as to have one reading per frame. This dataset features 300,000 high quality egocentric frames, each with a timestamp and its relative sensor reading. The dataset is fully annotated with six different motion classes: *biking*, *jogging*, *running*, *standing*, *walking* and *wandering*. The dataset is composed of videos captured in unconstrained real environments such as outdoor and indoor locations, captured by different people spanning various ages and in significantly different timespan featuring both night and day illumination conditions. Currently, there is no record of a publicly available egocentric video collection provided with accelerometer and gyroscope readings. Hence, being EGO-SENSORS the first significant dataset tackling with this problem, we publicly release it along with both the code[1].

[1]http://imagelab.unimore.it/sensors

	Linear SVM			
	Early Fusion		Our Approach	
Features	**Acc.**	**Std**	**Acc.**	**Std**
Blur + Optical Flow	0.724	0.017	0.741	0.015
A + G	0.762	0.054	0.921	0.009
Blur + A + G	0.792	0.059	0.909	0.009
Optical Flow + A + G	0.837	0.045	0.937	0.009
Blur + Opt. Flow + A + G	0.864	0.039	0.939	0.008

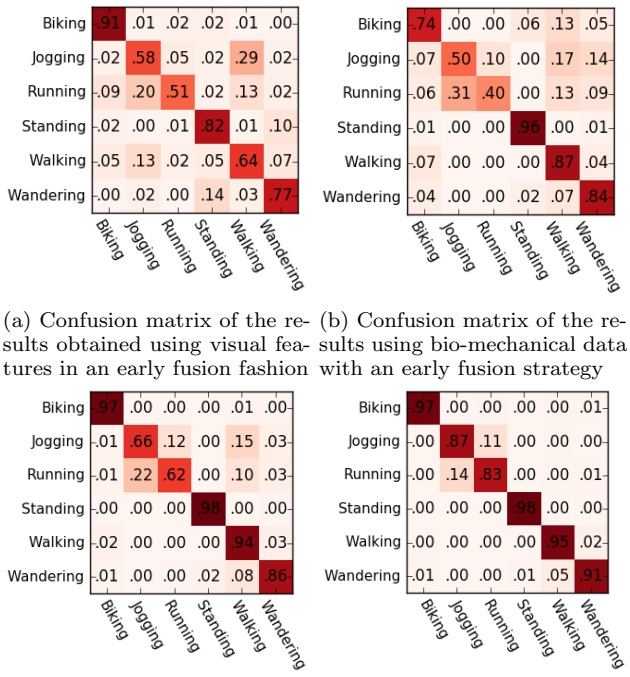
Table 1: Comparison of fusion strategies for different combinations of visual and bio-mechanical features. The first column describes the type of feature used, where A: Accelerometer, G: Gyroscope.

The experiments are conducted in an incremental way: first, we evaluate the contribution of different sets of features individually and then, the performance resulting from their combination is evaluated. To highlight the contribution of our data fusion strategy, results are reported with and without the use of our covariance descriptor, where the latter concatenates the different feature vectors (commonly referred to as early fusion). The features considered are: blur, optical flow, gyroscope and accelerometer information. Table 1 reports the results of the described evaluation in terms of classification accuracy and standard deviation. These metrics are obtained from a cross validation with a $80-20$ ratio between training and testing data where each test is repeated 100 times. All the experiments have been performed fixing the window sizes to $N = 4$ and $M = 40$, values identified with preliminary tests.

The table reports the contributions of different combination of visual and bio-mechanical descriptors. It can be noticed how the combination of different sensors can outperform their stand-alone usage in both fusion strategies. As expected, due to the unconstrained experimental setting solely adopting visual features results in poor performance. In fact, the table shows that the final feature vector that combines both visual and sensor data achieves the best results. It also shows that our solution, based on a covariance mapping, largely outperforms the standard early fusion strategy in all types of feature combinations. For example comparing the result of the final descriptor, we can observe that the propose solution improves over the early fusion approach by more than 8%.

For a more detailed analysis of the motion segmentation problem, we refer to the confusion matrices reported in Figure 3. In particular, Figure 3a reports the classification results using visual features in an early fusion fashion similar to what proposed for ego-motion segmentation by Lu *et al.* [10]. It can be seen how in this situation the class *walking* has a low accuracy rating due to the low discrimative capability of visual information to predict this particular class. On the other hand, Figure 3b reports results obtained using solely bio-mechanical features and shows a reduced risk of misclassifiying *walking* thanks to motion classification capabilities inherent of inertial sensors. Furthermore, the opposite situation occurs when classifying *biking* sequences, where visual information has better performance due to the fact that part of the bicycle is often visible in the scene. This shows the complementarity of the two information sources.

Considering the combination of features obtained from both visual and bio-mechanical sources, Figures 3c and 3d report the results obtained by using an early fusion approach



(a) Confusion matrix of the results obtained using visual features in an early fusion fashion

(b) Confusion matrix of the results using bio-mechanical data with an early fusion strategy

(c) Confusion matrix obtained by merging visual and bio-mechanical features through early fusion

(d) Confusion matrix obtained by merging visual and bio-mechanical features through our covariance descriptor

Figure 3: Confusion matrices of different features

(3c) and our covariance descriptor (3d). The Figures show how, given our descriptors' ability to capture the correlation between heterogeneous features, it can significantly reduce the error in the most ambiguous classes. For example, while the *jogging* class is recognized with a 66% accuracy using an early fusion strategy, our covariance descriptor can achieve a relative improvement of 32%.

4. CONCLUSION

In this paper we presented a thorough analysis of the multimodal acquisition capabilities of new egocentric wearable devices. We demonstrated that the main feature traditional computer vision focuses on, while providing a useful insight on the motion patterns present in the scene, can greatly benefit from the additional information resulting from bio-mechanical sensors. In particular, we design a novel data fusion strategy based on the projected covariance descriptor, which can better capture relations between visual and inertial information. Experimental results on a new and unconstrained egocentric dataset show that, when compared to the traditional early fusion strategy, our solution can improve the overall performance and significantly lower the error in the most ambiguous motion classes. Furthermore, due to the novelty of the task at hand, this work contributes to further research by releasing a real world unconstrained dataset annotated with motion activities and featuring both high quality video and bio-mechanical sensor readings.

5. REFERENCES

[1] S. Alletto, G. Serra, S. Calderara, F. Solera, and R. Cucchiara. From ego to nos-vision: Detecting social

relationships in first-person views. In *Proc. of Egocentric (First-Person) Vision Workshop, CVPRW*, 2014.

[2] A. Betancourt, P. Morerio, C. Regazzoni, and M. Rauterberg. The evolution of first person vision methods: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1–1, 2015.

[3] N. Brunetto, S. Salti, N. Fioraio, T. Cavallari, and L. Stefano. Fusion of inertial and visual measurements for rgb-d slam on mobile devices. In *Proc. of the IEEE International Conference on Computer Vision Workshops*, 2015.

[4] F. Crété-Roffet, T. Dolmiere, P. Ladret, M. Nicolas, et al. The blur effect: Perception and estimation with a new no-reference perceptual blur metric. In *Proc. of SPIE*, 2007.

[5] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis*, pages 363–370. Springer, 2003.

[6] A. Fathi, A. Farhadi, and J. Rehg. Understanding egocentric activities. In *Proc. of ICCV*, 2011.

[7] J. Hernandez, Y. Li, J. M. Rehg, and R. W. Picard. Bioglass: Physiological parameter estimation using a head-mounted wearable device. In *Proc. of International Conference on Wireless Mobile Communication and Healthcare*, 2014.

[8] K. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *Proc. of CVPR*, 2011.

[9] Q. Li, J. A. Stankovic, M. A. Hanson, A. T. Barth, J. Lach, and G. Zhou. Accurate, fast fall detection using gyroscopes and accelerometer-derived posture information. In *Proc. of International Workshop on Wearable and Implantable Body Sensor Networks*, 2009.

[10] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Proc. of CVPR*, 2013.

[11] S. Martelli, D. Tosato, M. Farenzena, M. Cristani, and V. Murino. An FPGA-based Classification Architecture on Riemannian Manifolds. In *Proc. of DEXA Workshops*, 2010.

[12] Y. Poleg, C. Arora, and S. Peleg. Temporal segmentation of egocentric videos. In *Proc. of CVPR*, 2014.

[13] S. D. Richard W. DeVaul. Real-time motion classification for wearable computing applications. In *Technical report, MIT Media Laboratory*, 2001.

[14] O. Tuzel, F. Porikli, and P. Meer. Pedestrian Detection via Classification on Riemannian Manifolds. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 30(10):1713–1727, 2008.