Human Activity Recognition Using a Dynamic Texture Based Method

Vili Kellokumpu, Guoying Zhao and Matti Pietikäinen
Machine Vision Group
University of Oulu, P.O. Box 4500, Finland
{kello,gyzhao,mkp}@ee.oulu.fi

Abstract

We present a novel approach for human activity recognition. The method uses dynamic texture descriptors to describe human movements in a spatiotemporal way. The same features are also used for human detection, which makes our whole approach computationally simple. Following recent trends in computer vision research, our method works on image data rather than silhouettes. We test our method on a publicly available dataset and compare our result to the state of the art methods.

1 Introduction

Human activity recognition has become an important research topic in computer vision in recent years. It has gained a lot of attention because of its important application domains like video indexing, surveillance, human computer interaction, sport video analysis, intelligent environments etc. All these application domains do have their own demands, but in general, algorithms must be able to detect and recognize various activities in real time. Also as people look different and move differently, the designed algorithms must be able to handle variations in performing activities and handle various kinds of environments.

Many approaches for human activity recognition have been proposed in the literature [4, 12]. Recently there has been a lot of attention towards analysing human motions in spatiotemporal space instead of analysing each frame of the data separately.

Blank et al. [1] used silhouettes to construct a space time volume and used the properties of the solution to the Poisson equation for activity recognition. Ke et al. [7] build a cascade of filters based on volumetric features to detect and recognize human actions. Shechtman and Irani [19] used a correlation based method in 3d whereas Kobyashi and Otsu [10] used Cubic Higher-order Local Autocorrelation to describe human movements.

Interest point based methods that have been quite popular in object recognition have also found their way to activity recognition. Laptev et al. [11] extended the Harris detector into space time interest points and detected local structures that have significant local variation in both space and time. The representation was later applied to human action recognition using SVM [17]. Dollár et al. [3] described interest points with cuboids, whereas Niebles and Fei-Fei [13] used a collection of spatial and spatial temporal features extracted in static and dynamic interest points.

BMVC 2008 doi:10.5244/C.22.88

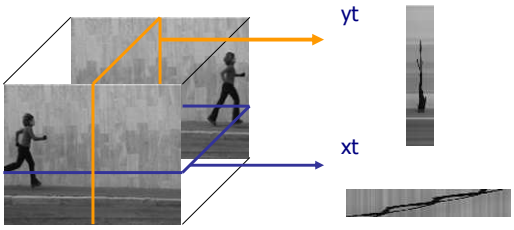


Figure 1. Illustration of a person running and the corresponding *xt* and *yt* planes from a single row and column. The different frames correspond to the *xy* planes.

Kellokumpu et al. [7] used a texture description to characterize Motion History Images and showed that a collection of local features can form a very robust description of human movements. We build on this idea and extend the idea of using histograms of local features into a spatiotemporal space. Furthermore, following recent trends in computer vision, we propose a method that is designed to work with image data rather than silhouettes. The method is based on using a dynamic texture descriptor, Local Binary Patterns from Three Orthogonal Planes (LBP-TOP), to represent human movements. The LBP-TOP features have successfully been used for facial expression recognition [22]. Niyogi and Adelson [14] proposed the use of *xt* slices for detecting contours of walking people. We propose a method for human detection that uses the LBP-TOP features (the same features we use for human motion description), making the combined approach computationally simple.

The rest of the paper is organized as follows. Section 2 introduces the dynamic texture descriptors. Section 3 describes their application to human detection and activity recognition. We show experimental results in Section 4 and conclude in Section 5.

2 Dynamic Texture Descriptors – LBP-TOP

LBP operator [15] describes local texture pattern with a binary code, which is obtained by thresholding a neighborhood of pixels with the gray value of its center pixel. An image texture can be described with a histogram of the LBP binary codes. LBP is a gray scale invariant texture measure and it is computationally very simple which makes it attractive for many kinds of applications. The LBP operator was extended to a dynamic texture operator by Zhao and Pietikäinen [22], who proposed to form their dynamic LBP description from three orthogonal planes (LBP-TOP) of a space time volume. Figure 1 shows the spatiotemporal volume of a person running from left to right. It also illustrates the resulting *xt* and *yt* planes from a single row of and column of the volume as well as the first and last *xy* planes that are the frames themselves. The LBP-TOP description is formed by calculating the LBP features from the planes and concatenating the histograms.

The original LBP operator was based on a circular sampling pattern but different neighbourhoods can also be used. Zhao and Pietikäinen proposed to use elliptic sampling for the *xt* and *yt* planes:

$$LBP\ (x_c, y_c) = \sum_{p}^{P-1} s(g_p - g_c) 2^P, \qquad s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}, \qquad (1)$$

where g_c is the gray value of the center pixel (x_c, y_c) and g_p are the gray values at the P sampling points: $(x_c\text{-}R_x\sin(2\pi p/P_{xt}), y_c, t_c\text{-}R_t\cos(2\pi p/P_{xt}))$ for xt plane and similarly $(x_c, y_c\text{-}R_y\sin(2\pi p/P_{yt}), t_c\text{-}R_t\cos(2\pi p/P_{yt}))$ for yt plane. R_d is the radius of the ellipse to direction of the axis d (x, y or t). As the xy encodes only the appearance, i.e., both axes have the same meaning, circular sampling is suitable. The values g_p for points that do not fall on pixels are estimated using bilinear interpolation.

In this work we consider only the usage of the temporal planes, namely the xt and yt planes. The reason for this is the variability in the appearance of humans and different environments. The xy plane contains a lot of useful appearance information but it should be noted that the temporal planes do also encode some of the low level appearance information.

3 Dynamic Texture Method for Human Motion Description

In this section we introduce a novel approach for human activity recognition. We use the LBP-TOP descriptors to locate a bounding volume of human in xyt space and then use the same features for describing human movements. Finally the temporal development of the features is modelled using Hidden Markov Models (HMMs).

3.1 Human Detection

Many approaches to human activity recognition rely on background subtraction for extracting the location and shape of people in video sequences. As the background subtraction is the first stage of processing in many human activity recognition systems, it has a huge effect on the overall performance of such systems. Also, many background subtraction methods are computationally expensive and memory demanding. This limits their possible usage in systems requiring processing at video rate.

We tackle the problem of computation cost by using the same features for both human detection and activity recognition. Usually background subtraction is done by modelling the pixel color and intensities [9, 20]. A different kind of approach was presented by Heikkilä and Pietikäinen [5] who introduced a region based method that uses LBP features from a local neighbourhood. They performed the subtraction by extracting the features in each frame. Unlike their work, we do not consider the image plane itself but instead the temporal planes xt and yt.

We adopt the idea of codebooks [9] in our approach and represent each local neighbourhood with a set of codes C. As observed by Heikkilä and Pietikäinen [5], the thresholding operation in the LBP feature extraction can be vulnerable to noise when pixel values of a neighborhood are close to one another. Therefore a bias is assigned to the center pixel which means that the term $s(g_p - g_c)$ in Eq. (1) is replaced with the term $s(g_p - g_c + a)$. Thus, our background model consists of codebook C and the bias a for each pixel for both the temporal planes.

We process the incoming data in overlapping volumes of duration Δt that is defined by R_t, i.e., $\Delta t = 2R_t+1$. Each volume has a center frame that forms the center pixels for the feature calculation and each frame acts as a center frame on its turn. If the observed LBP code of a pixel neighborhood of the input volume does not match the codes in the

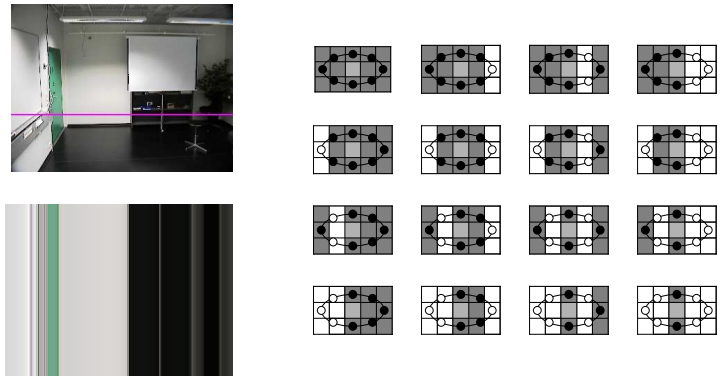


Figure 2. Illustration of LBP patterns that represent no motion. The two images on the left illustrate a state a static scene containing no motion and a resulting *xt* plane. The bit patterns illustrate the resulting codes that do not describe any motion. Consider nearest neighbor interpolation for the simplicity of the illustration and also note the bias on the center pixel.

corresponding codebook, the neighbourhood around the center pixel is determined to belong to an object. The result from *xt* and *yt* planes can be combined using the logical *and* operator. With this method, we can extract the bounding volume (3D equivalent to a bounding box in 2D) of a human in each space time volume.

For the current application, the detection part is not adaptive. The method is capable of extracting the rough human location and this is enough for our system. The proposed detection method can be extended to be adaptive to changes in the background and that topic is under investigation though out of the scope of this paper. We present preliminary experimental results in Section 4.

3.2 Activity Description

The previously introduced dynamic LBP features are used for human activity description and the input for activity recognition are the dynamic LBP features calculated from the detected human *xyt* volumes. However, a couple of points need addressing to enhance the performance of the features.

As we do not use silhouette data but rather an approximated bounding volume that contains the human, our input also contains some background information as well as the static human body parts. The appearance of these regions in the volume only depends on the structure of the scene and the clothing of the person and does not contain any useful information for motion description. Consider the images illustrated in Figure 2. Static parts of the images produce stripe like patterns for the *xt* and *yt* planes. As we wish to obtain a motion description, we can define (or learn by observing static scenes) those stripe patterns in the LBP representation and remove the corresponding bins from the histogram. The stripe patterns are always the same for a given LBP kernel only their relative appearance frequency depends on the scene structure. Figure 2 also illustrates these LBP codes for an eight point neighbourhood. Cutting off these bins reduces the histogram length for an eight point neighbourhood into 240 bins instead of 256, but more importantly, it also improves the motion description.

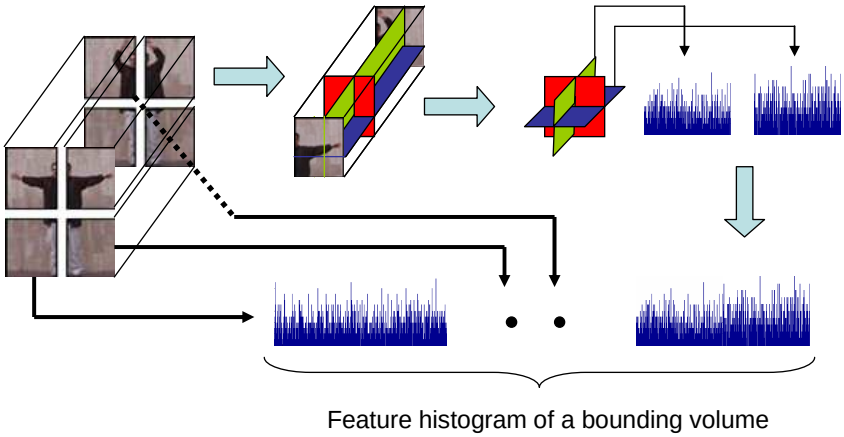


Figure. 3. Illustration of the formation of the feature histogram from a bounding volume.

The dynamic LBP features calculated over the whole bounding volume area encode the local properties of the movements without any information about their local or temporal locations. For this reason we partition the volume into subvolumes and form the feature histogram by concatenating the subvolume histograms. Using the subvolume representation we encode the motion on three different levels: pixel-level (single bins in the histogram), region-level (subvolume histogram) and global-level (concatenated subvolume histograms).

To obtain a rough spatial definition of human movements, we divide the detected bounding volume through its center point into four regions. This division roughly separates the hands and legs of the person in most viewpoints when the Δt is small or the person does not move much. Using more blocks would of course allow a more detailed description but would also produce more local histograms and make the whole histogram too long. Using too many blocks could also make the system too sensitive for stylistic variation of performing activities.

The subvolume division and the formation of our feature histogram are illustrated in Figure 3. All the subvolume histograms are concatenated and the resulting histogram is normalized by setting its sum equal to one.

3.3 Hidden Markov Model

As described earlier, we extract the dynamic features from a space time volume of short duration. We then model the development of our features using HMM. Our models are briefly described next but see tutorial [16] for more details on HMMs. In our approach a HMM that has N states $\mathbf{Q}=\{q_1,q_2,...,q_N\}$ is defined with the triplet $\lambda = (\mathbf{A},\boldsymbol{\pi},\mathbf{H})$, where $\mathbf{A}$ is the NxN state transition matrix, $\boldsymbol{\pi}$ is the initial state distribution vector and the $\mathbf{H}$ is the set of output histograms.

The probability of observing an LBP histogram h_{obs} is the texture similarity between the observation and the model histograms. Histogram intersection was chosen as the similarity measure as it satisfies the probabilistic constraints. Thus, the probability of observing h_{obs} in state i at time t is given as:

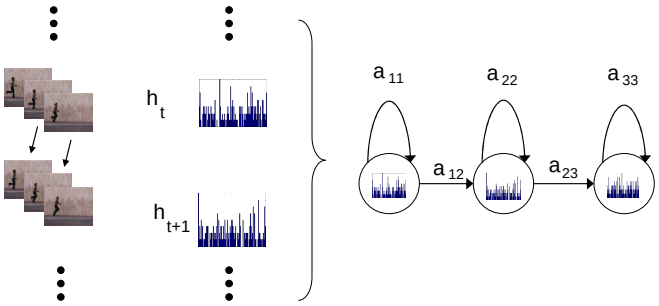


Figure 4. Illustration of the temporal development of the feature histograms and a simple HMM. This example shows a 3 state left-to-right HMM.

$$P(h_{obs} \mid s_t = q_i) = \sum \min(h_{obs}, h_i), \tag{2}$$

where s_t is the state at time step t, and h_i is the observation histogram in state i. The summation is done over the bins.

Figure 4 illustrates how the features are calculated as time evolves and a simple left-to-right HMM. We can use different kind of model topologies to model different kind of movements. Circular models are suitable for modelling repetitious movements like walking and running, whereas left-to-right models are suitable for movements like bending for example.

HMMs can be used for activity classification by training a HMM for each action class. A new observed unknown feature sequence $\mathbf{H}_{obs}=\{h_{obs1},h_{obs2},\ldots h_{obsT}\}$ can be classified as belonging to the class of the model that maximizes $P(\mathbf{H}_{obs}|\ \lambda)$, the probability of observing $\mathbf{H}_{obs}$ from the model λ. The model training is done using EM algorithm and the calculation of model probabilities can be done using forward algorithm.

4 Experiments

We demonstrate the performance of our method by experimenting with the Weizmann dataset [1]. The dataset has become a popular benchmark database [1, 2, 6, 8, 13, 18, 21] so we can directly compare our results to others reported in the literature.

The dataset consists of 10 different activities performed by 9 different persons. Figure 5 illustrates the activities. In the following subsections we show experimental results on human detection, feature analysis and human activity classification.

4.1 Detection

To get our background model we need to learn the codebook and bias for each pixel on two spatiotemporal planes. We train our model with the frames where there is no subject in the space time volumes.

In the background model training we calculate the codebook for each pixel with all bias values between a_{min} and a_{max}, and we choose the codebook (and the corresponding

Figure 5. Illustration of the movement classes in the Weizmann database. Starting from the upper left corner the movements are: Bending, Jumping jack, Jumping, Jumping in place ('Pjump'), Gallop sideways ('Side'), Running, Skipping, Walking, Wave one hand ('Wave1') and Wave two hands ('Wave2')

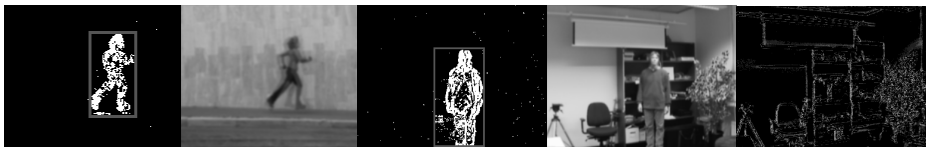


Figure 6. Illustration of the human detection performance. The last image on the right illustrates the bias for *yt* plane feature calculation for the scene viewed next to it. The binary result images illustrate the center pixels of the input volume that have been determined to not belong to background and the detected bounding volume.

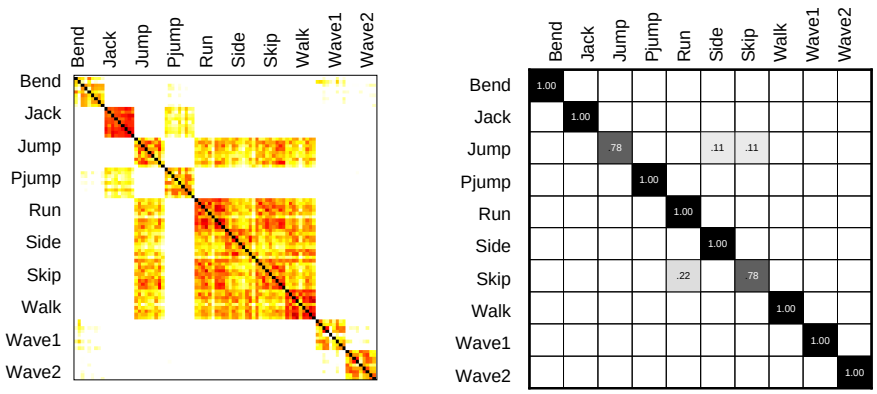
a) with the smallest number of codes. If codebook size is the same with multiple bias values, we choose the one with smallest bias. In our experiments we used a_{min}=3, a_{max}=8 and eight point neighbourhoods with radii R_x=1, R_y=2 and R_t=1 which means Δt=3. Figure 6 gives examples on the performance.

It should be noted that the learning method is preliminary and not optimal. But as mentioned earlier, the development of the detection part is out of the scope of this paper and under current work.

4.2 Feature Analysis and Activity Classification

First we want to illustrate how the proposed features themselves can describe the characteristics of different movements. We first calculated the LBP-TOP features for the dataset and for each movement we summed the histograms over the duration of the activity and normalized the histogram sum into one. This representation of the movements does not contain any temporal information.

Result of feature analysis is illustrated in Figure 7a where each row and column represent the similarity of one sample to all other samples. Histogram intersection was used as a similarity measure. The parameters used for the illustration are R_t=1, R_x=1, R_y=2 and P_{xt}=P_{yt}=8. It can clearly bee seen that even without any temporal information the features form clusters and different movements are somewhat separable.



(a) (b)

Figure 7. (a) Illustration of similarity of movements in the database, darker regions show higher similarity (b) the classification result.

We performed the activity classification experiments on the Weizmann dataset using HMM modelling and leave one out cross validation. The HMM topologies were set to circular for all cyclic activities and left-to-right models were used in other cases. We used HMMs with seven states and were able to classify 95.6% of the examples correctly using the same parameters as in the previous Subsection. Figure 7b shows the confusion matrix of the classification.

Results achieved by others on this database are summarized in Table 1. From the image based approaches Boiman and Irani [2] report the best overall recognition result, but their test set does not include the *Skipping* class. It is easy to see from the confusion matrix in Figure 7b that this extra class causes all but one of the mistakes we make in the test set. We also run the test without the skipping class and were able to classify 98.7% of the movements correctly. To our knowledge, our method gives the best results on the database when image data is used as an input and is also very competitive against approaches that are based on silhouette data.

Table 1. Results reported in the literature for the Weizmann database. The columns represent the reference, input data type, number of activity classes, number of sequences and finally the classification result

reference	input	act	seq	result
Our method	**image data**	**10 (9)**	**90 (81)**	**95.6% (98.7%)**
Scovanner et al. 2007 [18]	image data	10	92	82.6%
Boiman and Irani 2006 [2]	image data	9	81	97.5%
Niebles et al 2007 [13]	image data	9	83	72.8%
Kellokumpu et al 2008 [8]	silhouettes	10	90	97.8%
Wang and Suter 2007 [21]	silhouettes	10	90	97.8%
Ikizler and Duygulu 2007 [6]	silhouettes	9	81	100.0%

5 Conclusions and Future Work

We have proposed a novel dynamic texture based method for human activity recognition. We extract LBP-TOP features in spatiotemporal space and use them to detect human bounding volumes and to describe human movements. The method is computationally simple and utilizes image data rather than silhouettes, which makes it a suitable method for many applications. We have shown that our preliminary detection method can find human regions in spatiotemporal data and we show excellent results on human activity classification on a popular benchmark database.

As future work, we plan to develop the detection part and make it more accurate and adaptive to changes in the background. Also, as the *xy* plane contains much useful information we intend to examine how the data from the *xy* plane could be efficiently fused into the description.

Acknowledgement

The research was sponsored by the Graduate School in Electronics, Telecommunication and Automation (GETA) and the Academy of Finland.

References

[1] Blank M., Gorelick L., Shechtman E., Irani M. & Basri R. (2005) Actions as Space-Time Shapes. In Proc. ICCV, pp. 1395 – 1402.

[2] Boiman O. & Irani M. (2006) Similarity by Composition. In Proc. Neural Information Processing Systems (NIPS).

[3] Dollár P., Rabaud V., Cottrell G. & Belongie S. (2005) Behavior Recognition via Sparse Spatio-Temporal Features. In VS-PETS Workshop.

[4] Gavrila D. M. (1999) The Visual Analysis of Human Movement: A Survey. In CVIU, vol. 73, no. 3, pp. 82 – 98.

[5] Heikkilä M. & Pietikäinen M. (2006) A Texture-based Method for Modeling the Background and Detecting Moving Objects. In PAMI, vol. 28, no. 4, pp. 657 – 662.

[6] Ikizler N. & Duygulu P. (2007) Human Action Recognition Using Distribution of Oriented Rectangular Patches. In ICCV workshop on Human Motion Understanding, Modeling, Capture and Animation.

[7] Ke Y., Sukthankar R. & Hebert M. (2005) Efficient Visual Event Detection Using Volumetric Features. In Proc. ICCV, pp. 165 – 173.

[8] Kellokumpu V., Zhao G. & Pietikäinen M. (2008) Texture Based Description of Movements for Activity Analysis. In Proc. VISAPP, vol. 1, pp. 206 – 213.

[9] Kim K., Chalidabhongse T. H., Harwood D. & Davis L. (2004) Background Modeling and Subtraction by Codebook Construction. In Proc. ICIP, vol. 5, pp. 3061 – 3064.

[10] Kobyashi T. & Otsu N. (2004) Action and Simultaneous Multiple-Person Identification Using Cubic Higher-order Auto-Correlation. In Proc. ICPR, vol. 4, pp. 741 – 744.

[11] Laptev I. & Lindeberg T. (2003) Space-time Interest Points. In Proc ICCV, vol.1, pp. 432 – 439.

[12] Moeslund T. B., Hilton A. & Krüger V. (2006) A Survey of Advances in Vision-Based Human Motion Capture and Analysis. In CVIU, vol.104, issues 2-3, pp. 90 – 126.

[13] Niebles J. C. & Fei-Fei L. (2007) A Hierarchical Model of Shape and Appearance for Human Action Classification. In Proc. CVPR, 8p.

[14] Niyogi S. A. & Adelson E. H (1994) Analysing and Recognizing Walking Figures in XYT. In proc CVPR, pp. 469 – 474.

[15] Ojala T., Pietikäinen M. & Mäenpää T. (2002) Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. In PAMI, vol. 24, no. 7, pp. 971 – 987.

[16] Rabiner L. R. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, vol. 77, no. 2, pp. 257 – 285.

[17] Schuldt C., Laptev I. & Caputo B. (2004) Recognizing Human Actions: A Local SVM Approach. In Proc ICPR, pp. 32 – 36.

[18] Scovanner P., Ali S. & Shah M. (2007) A 3-Dimensional SIFT Descriptor and Its Application to Action Recognition. In Proc ACM Multimedia, pp. 357 – 360.

[19] Shechtman E. & Irani M. (2005) Space-Time Behavior Based Correlation. In Proc. CVPR, vol. 1, pp. 405 – 412.

[20] Stauffer C. & Grimson W. E. L. (1999) Adaptive Background Mixture Models for Real-Time Tracking. In Proc CVPR, vol. 2, pp. 246 – 252.

[21] Wang L. & Suter D. (2007) Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model. In Proc CVPR, 8p.

[22] Zhao G. & Pietikäinen M. (2007) Dynamic Texture Recognition Using Local Binary Patterns With an Application to Facial Expressions. In PAMI, vol. 29, no. 6, pp. 915 – 928.