Waving detection using the local temporal consistency of flow-based features for real-time applications*

Plinio Moreno, Alexandre Bernardino, and José Santos-Victor
{plinio, alex, jasv}@isr.ist.utl.pt

Instituto Superior Técnico & Instituto de Sistemas e Robótica
1049-001 Lisboa - Portugal

Abstract. We present a method to detect people waving using video streams from a fixed camera system. Waving is a natural means of calling for attention and can be used by citizens to signal emergency events or abnormal situations in future automated surveillance systems. Our method is based on training a supervised classifier using a temporal boosting method based on optical flow-derived features. The base algorithm shows a low false positive rate and if further improves through the definition of a minimum time for the duration of the waving event. The classifier generalizes well to scenarios very different from where it was trained. We show that a system trained indoors with high resolution and frontal postures can operate successfully, in real-time, in an outdoor scenario with large scale differences and arbitrary postures.

1 Introduction

Surveillance systems are becoming more and more frequent in urban areas and large public facilities (airports, shopping malls, stadiums). The number of installed cameras tends to grow as public security concerns increase. The utilization of networked robots and camera systems is also being investigated in international research projects [1] and may set the pace for future urban infrastructures. However, the security level has not been growing in proportion to the number of deployed cameras. Detection of security threats is done mostly by human operators that cannot deal with the huge amount of information that streams from the video sources. Even though some automated video surveillance systems have been proposed to detect some classes of events (like left luggage [2] and people fighting [3]) the number and the nature of possible security threats makes hard to develop a completely automated system. Our idea goes in the direction whereby citizens can help the surveillance system by signaling emergency, dangerous or suspicious situations with a universal alerting gesture: waving. As nowadays people dial emergency phone numbers to call for help, in the future they may just have to wave at any location covered by a surveillance system. Within this paradigm, we have been working in automatic and robust detection of waving events and this paper describes the current state of our research.

* Research partly funded by the FCT Programa Operacional Sociedade de Informação(POSI) in the frame of QCA III, and EU Project URUS (IST-045062)

1.1 Related Work

Detection on waving events can be framed in the current research on video based activity recognition. Several works have considered a general approach of action recognition, for instance aiming to distinguish among several different activities like walking, jogging, waving, running, boxing and clapping [4, 5]. The state-of-the-art research focus the limb tracking to model the human activities [6], an approach that is limited to high resolution targets and uncluttered environments [7]. In order to cope with cluttered environments, several works model activities using motion-based features [8, 3], shape-based features [9], space-time interest points [4] or a combination of some of the above features [10]. Although these works have achieved good recognition rates, the real-time performance is rarely mentioned by the authors, although the space-time "integral video" of [5] is driven by computational efficiency considerations.

1.2 Our Approach

In this paper we aim at a computational efficient representation of waving patterns by using motion-based features and a boosting classifier. We aim at performances comparable to the state-of-art but also able to run in real time in current video surveillance cameras. We exploit the constraints of fixed camera systems and develop a real-time waving detector that can be applied in indoors and outdoors scenarios. Our model of waving patterns relies on a qualitative representation of body parts' movements. Human activity is modeled using simple motion statistics information, not requiring the (time-consuming) pose reconstruction of parts of the human body. We use focus of attention (FOA) features [11] which compute optical flow statistics with respect to the target's centroid. In order to detect waving activities at every frame, a boosting algorithm uses labeled samples of FOA features in a binary problem: waving *vs* not waving. We use the Temporal Gentleboost algorithm [12] which improves boosting performance by adding a new parameter to the weak classifier: the (short-term) temporal support of the features. We improve the noise robustness of the boosting classification by defining a waving event, which imposes the occurrence of a minimum number of single-frame waving classifications in an suitably defined temporal window. The robustness of the waving model (FOA and GentleBoost) is tested on the KTH action database and compared to the state-of-the-art results.

The main requirement of the waving model proposed in this work is the previous segmentation and labeling of moving targets in the image. Due to real-time performance constraints, we adopt fast algorithms for segmentation and labeling. Since detection will be performed in a network of fixed cameras, the initial segmentation is provided by a background subtraction algorithm. We use the Lehigh Omnidirectional Tracking System (LOTS) method [13], which adapts the background by incorporating the current image with a small weight. After getting a new image, the background detection process generates a list of bounding boxes corresponding to connected foreground objects in the image. Then, the tracking algorithm performs data association between consecutive frames, using the distance between centroids of the bounding boxes. The user can select the data association algorithm, according to the desired performance: a fast nearest

neighbor or the more robust hungarian assignment [14]. In the image regions corresponding to the detected targets, we compute FOA features based on a dense optical flow algorithm [15]. The optical flow algorithm is based on a new metric for intensity matching, which removes noisy flow vectors with a low computational load.

We show in both indoors and outdoors datasets the robustness and generalization properties of the approach, attaining high frame rate performance (up to 20fps) and low false positive rate. In addition, the method is able to detect waving patterns in low resolution targets, which is frequently the case in cameras with wide field of view. In section 2, we describe the waving model in detail and evaluate its properties, then in section 3 we explain the real-time implementation, followed by the results in section 4 and conclusions in section 5.

2 Waving model

In this section we explain what image features and classification techniques are used to be able to detect waving gestures in a stream of video.

2.1 Focus Of Attention (FOA) features

FOA features encode the motion patterns of parts of the body with respect to its center [11]. This representation is based on the statistical distribution of the optical flow in the image region corresponding to the detected targets. We assume that the center of the bounding box corresponds rougly to the center of the person's body, and then the following computations are performed:

1. The mean value of the optical flow is computed around several angular directions with respect to the centroid of the target's segmented pixels. Particular gestures involve motion of body parts within a limited range of angles. For instance, the expected angular variation of legs during walking span a certain range $\Delta\theta$, as illustrated in the left part of Figure 2.1. A range $\Delta\theta$ can be seen as an receptive field tuned for the extraction of the movement of a particular part of the body.
2. For each angle, the optical flow vectors within the receptive field are pooled and projected on the radial and normal directions. The final motion representation is the concatenation of such projections for all angles (with an appropriate discretization). The right part of Figure 2.1 shows an example of the mean optical flow vector at the arm direction.

Different types of body movements will activate different receptive fields in different ways, forming characteristic patterns that represent basic movements like rising/putting down arms, bending, sitting, etc. The response of the receptive fields forming the FOA representation at each time will provide the information required to identify such basic movements.

2.2 Temporal Gentleboost

To train classifiers able to recognize waving patterns in images, we use a boosting algorithm. Boosting algorithms provides a framework to sequentially fit additive models

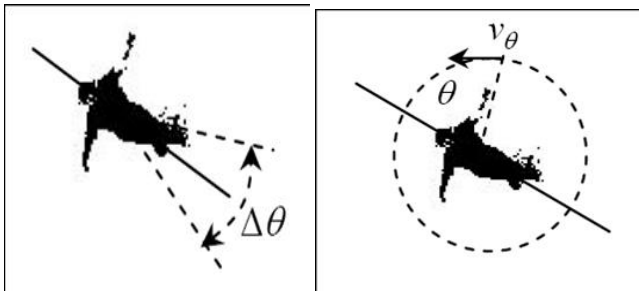

Fig. 1. Focus Of Attention examples.

in order to build a classifier:

$$H(x) = \text{sign}\left(\sum_{m=1}^{M} h_m(x)\right) \tag{1}$$

In the previous equation H is called *the strong classifier* and is obtained by the computing the sign of the sum of M weak classifiers h. Variable x denotes the vector of FOA features we want to classify. If $H(x) = 1$ we detect a positive example whereas if $H(x) = -1$ no detection is obtained from x. The training of such a classifier , h_m, is done by minimizing at each round m the weighted squared error $J = \sum_{i=1}^{N} w_i(y_i - h_m(x_i))^2$ with respect to the classifier parameters, where N is the number of training samples, y_i are the ground truth values (1 for detection and -1 for no detection) and $w_i = e^{-y_i h_m(x_i)}$ are weights. At each round, the weak classifier with lowest error is then added to the strong classifier and the data weights adapted, increasing the weight of the misclassified samples and decreasing correctly classified ones [16].

We use a particular class of boosting algorithm called GentleBoost [16] that commonly uses very simple functions, known as regression stumps, to implement the weak classifiers. Regression stumps have the form $h_m(x) = a\delta\left[x^{(f)} > \theta\right] + b\delta\left[x^{(f)} \leq \theta\right]$, where the scalar $x^{(f)}$ is the f^{th} entry of data sample x. Function δ is an indicator, i.e. δ[*condition*] is one if *condition* is *true* and zero otherwise. Regression stumps can be viewed as bifurcations on decision trees, where the indicator function sharply chooses branch a or b depending on threshold θ and feature $x^{(f)}$. To optimize the stump one must find the set of parameters $\{a, b, f, \theta\}$ that minimize J. A closed form exists to compute the optimal a and b, and the pair $\{f, \theta\}$ is found using exhaustive search [17].

A recent approach considers the temporal evolution of the features in the boosting algorithm, improving its noise robustness and performance [12]. That work models the temporal consistency of the features by parameterizing time in the weak classifiers. The Temporal Stumps compute the mean classification output of the regression stump, in a temporal window of size T,

$$h_m(x_i) = a\left(\frac{1}{T}\sum_{t=0}^{T-1}\delta\left[x_{i-t}^{f} > \theta\right]\right) + b\left(\frac{1}{T}\sum_{t=0}^{T-1}\delta\left[x_{i-t}^{f} \leq \theta\right]\right). \tag{2}$$

1. Given: $(x_1, y_1), \ldots, (x_N, y_N)$ where $x_i \in X$, $y_i \in Y = \{-1, +1\}$, set $H(x_i) := 0$, initialize the observation weights $w_i = 1/N$, $i = 1, 2, \ldots, N$
2. Repeat for $m = 1, \ldots, M$
   (a) Find the optimal weak classifier h_m over (x_i, y_i, w_i).
   (b) Update strong classifier $H(x_i) := H(x_i) + h_m^*(x_i)$
   (c) Update weights for examples $i = 1, 2, \ldots, N$, $w_i := w_i e^{-y_i h_m^*(x_i)}$

Fig. 2. Temporal Gentleboost algorithm.

The temporal weak classifier of Eq. 2 can be viewed as the classic regression stump with a different "indicator function". If $T = 1$ it becomes the original regression stump, and for $T > 1$ the indicator function changes. The new indicator functions are

$$\Delta_+^T(f, \theta, T) = \frac{1}{T} \sum_t^{T-1} \delta \left[x_{i-t}^f > \theta \right], \quad \Delta_-^T(f, \theta, T) = \frac{1}{T} \sum_t^{T-1} \delta \left[x_{i-t}^f \leq \theta \right], \qquad (3)$$

and compute the percentage of points above and below the threshold θ, in the temporal window T, for the feature number f. The indicator functions with temporal consistency in Eq. 3, can take any value in the interval $[0\,1]$, depending on the length of the temporal window used. For example, if $T = 2$ the functions can take 3 different values, $\Delta_+^T \in \{0,\ 1/2,\ 1\}$, if $T = 3$ can take four values, $\Delta_+^T \in \{0,\ 1/3,\ 2/3,\ 1\}$ and so on. The output of the new "indicator function", Δ, represents the confidence on the threshold selection to use the data with temporal support T. Thus, at each boosting round, we use a weighted confidence of both branches, instead of choosing only one branch.

Using the weak classifier with temporal consistency of Eq. 2 in the cost function it is possible to obtain closed form solutions for the parameters a and b that minimize the error J [12]. The optimal f,θ and T are obtained by exhaustive search. The learning algorithm shown in figure 2 is similar to GentleBoost, but optimizes the temporal stump of Eq. (2).

2.3 Waving event

The temporal boost algorithm described above improves the single frame classification of the waving activity using the short-term consistency of the FOA features. However, there are problems at the on-set and off-set of the waving gestures both in the generation of ground truth data and on the classification output. Also, some other gestures have short term similarity with waving actions but are of different nature. Thus, in order to reduce the false positive rates we require waving gestures to be persistent for a few frames otherwise they are discarded. This is in accordance with the human behavior as we usually wave for a long enough time if we want to make sure our sign is detected. We define the waving event as active when occurs a minimum number of single-frame waving classifications in a temporal window. In addition, the filtering property of the event definition can be adapted to different frame rates by selecting the value of the temporal window.

2.4 Evaluation of the waving model

The model presented in the previous sections exploits the characteristics of the motion of the waving activity and its temporal extent. In this section we evaluate the suitability of this model in the KTH action database [4], which has video sequences of six activities: walking, running, jogging, boxing, clapping and waving. We use a subset of this database in order to distinguish between waving and the negative samples (boxing and clapping). The negative samples selected have motion patterns very similar to those of the waving activity and the experimental results from previous works support this selection [4, 5, 18].

For this comparison we use the training and testing set of [4]. A user clicks in the first image of every sequence to provide the centroid of the targets. Then, the FOA features are computed in the entire image ($\Delta\theta = \pi/4$), using the dense optical flow of [15]. The final step is the supervised temporal boost learning followed by the single-frame classification and event classification. The accuracy of classifying every sequence correctly (i.e. classifying correctly the occurence of the waving event in the sequence,) is shown in Table 1.

Related work	Accuracy
Our method	**91.7%**
Niebles et al. [18]	93%
Ke et al. [10]	88%
Ke et al. [5]	91.7%
Schuldt et al. [4]	73.6%

Table 1. Accuracy of state-of-the-art methods in waving detection on the KTH action database. In our method, the temporal support of the Temporal boost algorithm is 25 frames (1s) and the classification of every sequence uses an event window size of 4s, considering a waving event if at least 60% of the single-frame classifications are positive in that sequence.

Our model for waving detection has a performance comparable to the state-of-the-art with the advantage of a very low computational load at detection time. We have implementation running in real time (20fps) on full sized images (640x480).

3 Real-time implementation

The robustness and real-time performance of the presented system partly rely on the employed target segmentation and tracking methods. In our case we use the LOTS background subtraction for segmentation [13] and distance based data association for tracking. In addition, the FOA features computation rely on the fast optical flow implementation of [15] which presents a good balance between speed and quality.

Like many segmentation systems, LOTS processing starts with a change-detection method based on background subtraction. The main difficulties of such approach lie in the fact that, even in controlled environments, the background undergoes a continual

change, mostly due to the existence of lighting variations and distractors (*i.e.,* clouds passing by, branches of trees moving with the wind). Target occlusion and interaction with the scene rises additional problems. To overcome these difficulties, the robust and fast algorithm described in [13] was implemented. The robustness towards lighting variations of the scene is achieved using adaptive background models and adaptive per-pixel thresholds. The use of multiple backgrounds and grouping pixels through quasi-connected-components (QCC) contribute to the robustness of the algorithm towards unwanted distractors.

The LOTS algorithm provides the bounding boxes of the regions of interest and their corresponding segmented pixels. The distance between the center points of two bounding boxes is the feature selected to do data association between consecutive frames. The user has two options for data association algorithms: (i) nearest neighbor or (ii) hungarian assignment. The nearest neighbor is the more efficient option, while the hungarian assigment minimizes the global cost of the assignments in polynomial time. The hungarian algorithm[1] works better than the nearest neighbor when the paths of two or more targets intersect each other. However, the computational load of the hungarian algorithm may be a problem when tracking a large number of targets (greater than 10).

In addition to the segmentation and labeling techniques, the computational load of the optical flow algorithm (dense) is crucial to attain high frame rates. We use the implementation of [15][2], an optical flow algorithm that introduces a new metric for intensity matching, based on the unequal matching (i.e. unequal number of pixels in the two images can be correspondent to each other). The optical flow used has a good balance between computational load and robustness to noise in the motion estimation [15]. The software was implemented in C++ using YARP libraries, using a P4(2.8GHz) PC. The frame rate of the waving detector varies according to the setup of the algorithms, as follows:

- **Frame rate: 20fps**. LOTS algorithm uses images of size 640×480, the optical flow uses images of size 160×120 and is computed only in the bounding boxes.
- **Frame rate: 10fps**. LOTS algorithm uses images of size 640×480, the optical flow uses images of size 320×240 and is computed only in the bounding boxes.

4 Experiments and results

The real-time implementation of the waving detector was developed specifically for the URUS project [1] and was trained and tested on different databases, considering two sequences for training and one sequence for testing. Figure 4 shows one sample of each data set, which contains several actions of the negative class (walking, pointing, and wandering). The training sequences have 4229 frames (2303 waving and 1926 not waving), and the testing sequence has 4600 frames (1355 waving and 3245 not waving). The FOA feature sampling is $\Delta\theta = \pi/4$. The support window of the Temporal boost algorithm is 20 frames. The event window size is 2s (20 frames), considering a waving event if at least 60% of the single-frame classifications are positive. Table 2 shows the robustness improvement obtained by the definition of the event in both data sets.

[1] obtained from http://www.mathworks.com/matlabcentral/fileexchange/6543

[2] http://www.cs.umd.edu/~ogale/download/code.html

	single-frame	Event
training set	92.01%	92.74%
testing set	85.95%	94.43%

Table 2. Waving detector accuracy on the sequences of Figure 4.

The definition of the event brings robustness to noisy classifications, improving up to 9% the accuracy of the results. In addition, the event window size can be adapted to different frame rates. Figure 4 shows examples of waving events detected correctly in the case of sequences grabbed at 10fps, in which the event window size is 1s.

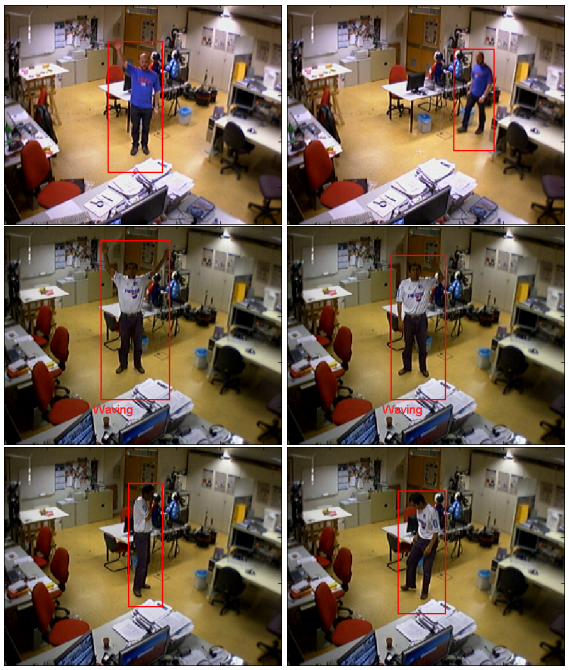

Fig. 3. Sample images of the indoors data set. In the first row, positive and negative samples of the training set. In the second row, samples of waving events correctly detected. In the third row, samples of the negative class (not waving) correctly detected.

Though we do not yet have performed a quantitative analysis of the outdoor results, we noticed that the classifier generalizes well to conditions very distinct from the ones on the training data, in terms of scale (trained with large targets but also detects small ones), lighting (trained indoors but also works outdoors), and posture (trained with frontal postures but also detects lateral postures).

Fig. 4. Sample images of waving events correctly detected in different scenarios. Notice the correct detection of the bottom images, where the subject is waving away from the camera.

5 Conclusions

We have addressed the real-time detection of waving gestures in fixed camera systems, showing its application in indoors and outdoors settings. The waving model extracts motion information of the targets using the statistics of optical flow features. Then the temporal boost algorithm learns to discriminate between waving and other patterns. In addition, the definition of a waving event by pooling the results of the classification result in a temporal window, adds robustness to the detection. The model presented is efficient and accurate, with performance comparable to the state-of-the-art approaches.

The adopted algorithms for segmentation, data association and optical flow computation have a low computational load, thus enabling the real-time execution of the waving detection algorithm. In future work, the addition of an efficient person detector should remove erroneous segmentations provided by the background subtraction algorithm. Also, a tracking algorithm with richer features will certainly increase the robustness of the waving detections.

References

1. Sanfeliu, A., Andrade-Cetto, J.: Ubiquitous networking robotics in urban settings. In: Workshop on Network Robot Systems. Toward Intelligent Robotic Systems Integrated with Environments. Proceedings of 2006 IEEE/RSJ International Conference on Intelligence Robots and Systems (IROS2006). (2006)

2. Ribeiro, P., Moreno, P., Santos-Victor, J.: Detecting luggage related behaviors using a new temporal boost algorithm. In: Proc. of PETS 2007 - 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, in conjunction with ICCV 2007 - 11th International Conference on Computer Vision, 2007. (2007)
3. Ribeiro, P., Santos-Victor, J.: Human activities recognition from video: modeling, feature selection and classification architecture. In: BMVC Workshop on HAREM. (2005)
4. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Volume 3. (2004) 32–36 Vol.3
5. Ke, Y., Sukthankar, R., Hebert, M.: Efficient visual event detection using volumetric features. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. Volume 1. (2005) 166–173 Vol. 1
6. Ikizler, N., Forsyth, D.: Searching video for complex activities with finite state models. In: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on. (2007) 1–8
7. Poppe, R.: Vision-based human motion analysis: An overview. Comput. Vis. Image Underst. **108**(1-2) (2007) 4–18
8. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision, Washington, DC, USA, IEEE Computer Society (2003) 726
9. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Trans. Pattern Anal. Mach. Intell. **29**(12) (2007) 2247–2253
10. Ke, Y., Sukthankar, R., Hebert, M.: Spatio-temporal shape and flow correlation for action recognition. In: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on. (2007) 1–8
11. Pla, F., Ribeiro, P.C., Santos-Victor, J., Bernardino, A.: Extracting motion features for visual human activity representation. In: Proceedings of the IbPRIA'05. (2005)
12. Ribeiro, P.C., Moreno, P., Santos-Victor, J.: Boosting with temporal consistent learners: An application to human activity recognition. In: Proc. of 3rd International Symposium on Visual Computing. (2007) 464–475
13. Boult, T.E., Micheals, R.J., Gao, X., Eckmann, M.: Into the woods: Visual surveillance of noncooperative and camouflaged targets in complex outdoor settings. Proceedings Of The IEEE **89**(10) (2001) 1382–1402
14. Ahuja, R., Magnanti, T., Orlin, J.: Network Flows. Prentice Hall (1993)
15. Ogale, A.S., Aloimonos, Y.: A roadmap to the integration of early visual modules. International Journal of Computer Vision **72**(1) (2007) 9–25
16. J. Friedman, T.H., Tibshirani, R.: Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Annals of Statistics **28**(2) (2000) 337407
17. Torralba, A., Murphy, K., Freeman, W.: Sharing visual features for multiclass and multiview object detection. IEEE Transactions On Pattern Analysis and Machine Intelligence **29**(5) (2007) 854–869
18. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision **79**(3) (2008) 299–318