An Attention Based Method For Motion Detection And Estimation

Shijie Zhang and Fred Stentiford
Department of Electronic and Electrical Engineering
University College London, Adastral Park Campus

Abstract

The demand for automated motion detection and object tracking systems has promoted considerable research activity in the field of computer vision. A novel approach to motion detection and estimation based on visual attention is proposed in the paper. Comparisons are made with the Berkeley MPEG-1 video analyzer [1]. Preliminary results show that the new method extracts more information about the moving object than that available in the MPEG encoding. In addition the method does not suffer from some of the inaccuracies inherent in the MPEG encoding.

1. Introduction and Background

The demand for automated motion detection and object tracking systems has promoted considerable research activity in the field of computer vision [2-6]. Bouthemy [2] proposed a novel probabilistic parameter-free method for detecting independently moving objects using the Helmholz principle. Optical flow fields were estimated without making assumptions on motion presence and allowed for possible illumination changes. The method imposes a requirement on the minimum size for the detected region. Also detection errors arise with small and low contrasted objects. Black and Jepson [3] proposed a method for optical flow estimation based on the motion of planar regions plus local deformations. The approach used brightness information for motion interpretation by using segmented regions of piecewise smooth brightness to hypothesize planar regions in the scene. The proposed method still has a problem dealing with small and fast moving objects. It is also computational expensive. Viola and Jones [4] presented a pedestrian detection system that integrated both image intensity (appearance) and motion information, which was the first approach that combined motion and appearance in a single model. The system works relatively fast and operates on low resolution images under difficult conditions such as rain and snow. However, it does not detect occluded or partial human figures. In [5] a method for motion detection based on a modified image subtraction approach was proposed to determine the contour point strings of moving objects. The proposed algorithm works well in real time and is stable for illumination changes. However, it is weak in areas where a contour appears in the background which corresponds to a part of the moving object in the input image. Also some of the contours in temporarily non-moving regions are neglected in memory so that small broken contours may appear. In [6] the least squares method was used for change detection. The proposed approach is efficient and successful on image sequences with low SNR and is robust to illumination changes. The biggest shortfall is that it can only cope with single object movements.

The use of visual attention (VA) methods [7-9] to define the foreground and background information in a static image for scene analysis has motivated this investigation. We propose in this paper that similar mechanisms may be applied to the detection of saliency in motion and thereby derive an estimate for that motion.

2. Motion VA Algorithm

Methods for identifying areas of static saliency are described in [8] and are based upon the premise that regions which are largely different to most of the other parts of the image will be salient and will be present in the foreground. Such decisions between foreground and background could be dependent upon features such as colour, shape, texture, or a combination. This concept has been extended into the time domain and is applied to sequences of video frames to detect salient motion. This approach is not dependent on a specific segmentation process but only upon the detection of anomalous movements.

In order to reduce computation candidate regions of motion are first detected by generating the intensity difference frame from adjacent frames and applying a threshold.

$$I = \{|r_2 - r_1| + |g_2 - g_1| + |b_2 - b_1|\}/3$$

Where parameters (r_1, g_1, b_1) & (r_2, g_2, b_2) represent the rgb colour values for frame 1 and 2. The intensity I is calculated by taking the average of the differences of rgb values between the two frames.

The candidate regions R_I in frame 1 are then identified where $|I| > T$ and T is a threshold.

Let a set of measurements $\boldsymbol{a}$ = *(r, g, b)* correspond to a location $\boldsymbol{x}$ = *(x, y)* in R_t

Define a function $\boldsymbol{F}$ such that $\boldsymbol{a} = \boldsymbol{F(x)}$ and let $\boldsymbol{x_0}$ be in R_t in frame t.

Consider a neighbourhood *G* of $\boldsymbol{x_0}$ within a window of radius ε where

$$\left\{\boldsymbol{x}_i' \in G \quad iff \quad \left|x_0 - x'\right| \le \varepsilon \; and \; \left|y_0 - y'\right| \le \varepsilon\right\}$$

Select a set of *m* random points S_x in *G* (known as the fork) where

$$S_x = \left\{\boldsymbol{x}_1', \boldsymbol{x}_2', ..., \boldsymbol{x}_m'\right\}$$

We also only consider forks which are constrained to contain pixels that mismatch each other. This means that forks will be selected in image regions possessing high or certainly non-zero VA scores, such as on edges or other salient features.

In this case there will be at least one pixel in the fork that differs by more than δ in one or more of its rgb values with one or more of the other pixels in the fork i.e.

$$\left|F_k(\boldsymbol{x}_i') - F_k(\boldsymbol{x}_j')\right| > \delta_k \quad for \; some \; i, j, k$$

Define the radius of the region within which fork comparisons will be made as *V* (the *view radius*)

Randomly select another location $\boldsymbol{y_0}$ in the adjacent frame R_{t+1} within a radius *V* of $\boldsymbol{x_0}$.

Define the fork $S_y = \left\{\boldsymbol{y}_1', \boldsymbol{y}_2', ..., \boldsymbol{y}_m'\right\}$ where $\boldsymbol{x_0} - \boldsymbol{x}_i' = \boldsymbol{y_0} - \boldsymbol{y}_i'$

and $\left|x_{0j} - x_{ij}'\right| \le R$

The fork centred on $\boldsymbol{x_0}$ is said to match that at $\boldsymbol{y_0}$ (S_x matches S_y) if

$$\left|F_j(\boldsymbol{x_0}) - F_j(\boldsymbol{y_0})\right| < \delta_j$$

and $\left|F_j(\boldsymbol{x}_i') - F_j(\boldsymbol{y}_i')\right| < \delta_j \quad j = r, g, b \quad i = 1,2,..., m.$

N attempts are made to find matches and the corresponding displacements are recorded as follows:

Let the *i*th fork S_{xi} match S_{yi}, where $S_{xi} = \left\{\boldsymbol{x}_{1i}', \boldsymbol{x}_{2i}', ..., \boldsymbol{x}_{mi}'\right\}$ and $S_{yi} = \left\{\boldsymbol{y}_{1i}', \boldsymbol{y}_{2i}', ..., \boldsymbol{y}_{mi}'\right\}$

Define the matching displacement as $\boldsymbol{\sigma}^{t+1} = \left(\sigma_p, \sigma_q\right)$ where

$$\sigma_p = \left|x_{0p} - y_{0p}\right|, \quad \sigma_q = \left|x_{0q} - y_{0q}\right|$$

and the cumulative displacements Δ and match counts Γ as

$$\left.\begin{aligned}\Delta(\boldsymbol{x}_j) &= \Delta(\boldsymbol{x}_j) + \boldsymbol{\sigma}_j \\ \Gamma(\boldsymbol{x}_j) &= \Gamma(\boldsymbol{x}_j) + 1\end{aligned}\right\} \; j = 1,..., N_1 < N$$

where N_1 is the total number of matching forks and *N* is the total number of matching attempts

The displacement $\overline{\boldsymbol{\sigma}}_{\boldsymbol{x_0}}^{t+1}$ corresponding to pixel $\boldsymbol{x_0}$ averaged over the matching forks is

$$\overline{\boldsymbol{\sigma}}_{\boldsymbol{x_0}}^{t+1} = \frac{\sum_{j=1}^{N_1} \Delta(\boldsymbol{x}_{ji})}{\sum_{j=1}^{N_1} \Gamma(\boldsymbol{x}_{ji})}$$

A similar calculation is carried out between R_t and R_{t-1} to produce $\overline{\boldsymbol{\sigma}}_{\boldsymbol{x_0}}^{t-1}$ and the estimated displacement of $\boldsymbol{x_0}$ is given by $\{\overline{\boldsymbol{\sigma}}_{\boldsymbol{x_0}}^{t+1} - \overline{\boldsymbol{\sigma}}_{\boldsymbol{x_0}}^{t-1}\}/2$. This estimate takes account of both trailing and leading edges of moving objects.

This is carried out for every pixel $\boldsymbol{x_0}$ in the candidate motion region R_t and M attempts are made to find an internally mismatching fork S_x.

3. Results and Discussion

A pair of frames from a traffic video was tested on with results shown in Figure 1. The intensity difference indicates the areas of candidate motion for subsequent analysis. Motion vectors (MVs) were calculated as above for the car region and plotted in Figure 2. Regions with no vector assigned are those where no internally mismatching forks could be found i.e. areas of similar colour. The processing took 40 seconds in Matlab 7.
The parameters of the experiment were $M = 100$, $N = 100$, $\varepsilon = 3$, $m = 7$, $R = (10,10)$, $\boldsymbol{\delta} = (40,40,40)$, $T = 40$.

frame 1 frame 2 intensity difference

Figure 1. Two adjacent frames and their intensity difference

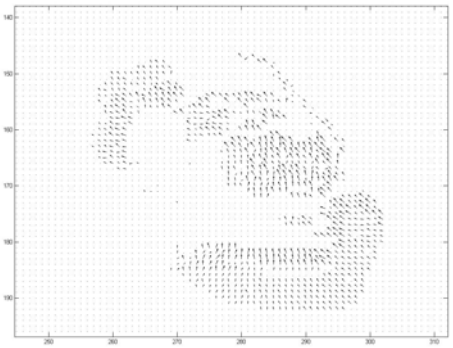
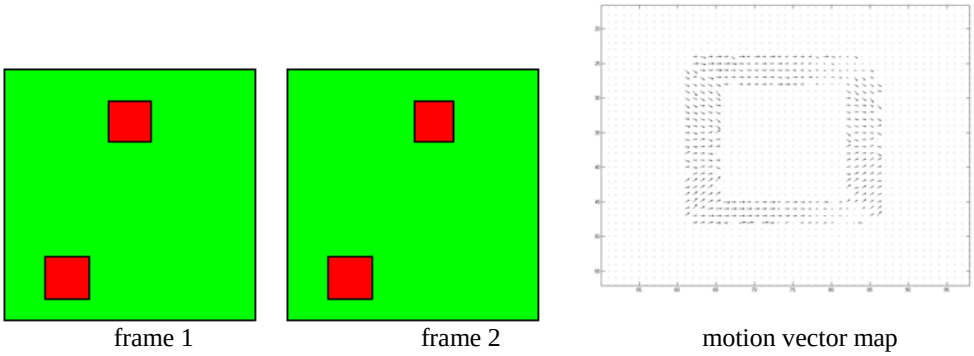


Figure 2. Motion vector map

It is possible to deduce information about the shape of the moving object from the motion vector map. Figure 3 shows two red squares in each frame, with the upper one moving from left to right and the other remaining static. A motion vector map for the moving square is shown alongside. The nature of the fork matching process means that there is a displacement component superimposed that points towards the centre of curvature along the boundary of the object. Motion vectors near corners therefore point towards the centre of the square in Figure 3, but all vectors possess a component in the direction of motion.



frame 1 frame 2 motion vector map

Figure 3. Synthetic data

Figure 4 illustrates a comparison between the motion vectors present in the MPEG encoding [1] and those derived from the motion VA algorithm. The motion vector map in Figure 2 was scaled from 352x288 pixels to 22x18 macroblocks with each block corresponding to a 16x16 pixel area in the original image. It can be seen that the

MPEG vectors are not as localised and several are spurious. In fact one of the vectors is in the wrong direction. Since MPEG encoding criteria are minimizing the number of bits being used and satisfying the constraints on frame types (e.g. no MVs in I-frames), this particular block was encoded differently in adjacent frames. This means that the MV for that block cannot be calculated by subtracting MVs in adjacent frames. In practice, MVs are essentially some combination of bits that the encoder can use to select between to reduce the bits required to encode a block. Having more options to search, i.e. more MVs to choose from, and more CPU time leads to more accurate image coding, but not necessarily more accurate MVs.

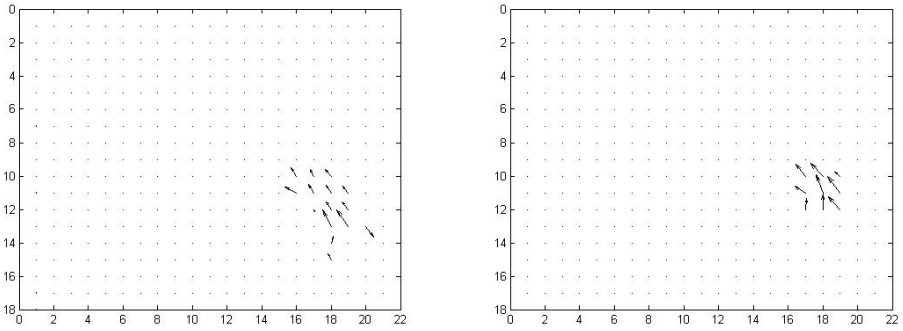


Figure 4. Motion vector maps extracted from MPEG-1 analyzer (left) and VA algorithm (right)

4. Conclusions

An attention based mechanism has been proposed for motion detection and estimation. The work indicates that there is potential for extracting shape information about the moving objects and also for overcoming some of the inherent inaccuracies in MPEG motion vectors.

Future work will be carried out to investigate using more accurate colour spaces (CIE standard), as well as addressing background motion and changes in illumination.

5. Acknowledgement

The project is sponsored by European Commission Framework Programme 6 Network of Excellence MUSCLE (Multimedia Understanding through Semantics, Computation and Learning) [11].

6 References

[1] Moving Picture Experts Group Homepage "http://mpeg.telecomitalialab.com" Berkeley Multimedia Research Center, http://bmrc.berkeley.edu

[2] T. Veit, F. Cao, and P. Bouthemy, “Probabilistic parameter-free motion detection,” CVPR, vol. 1, pp. 715-721, June 27-July 2, 2004.

[3] M.J. Black and A.D. Jepson, “Estimating optical flow in segmented images using variable-order parametric models with local deformations,” IEEE Trans. on PAMI, vol. 18, Issue 10, pp. 972-986, Oct. 1996.

[4] P. Viola, M.J. Jones, and D. Snow, “Detecting pedestrians using patterns of motion and appearance,” ICCV, vol. 2, pp. 734-741, Oct. 2003.

[5] M. Kellner and T. Hanning, “Motion detection based on contour strings,” ICIP, vol. 4, pp. 2599-2602, Oct. 2004.

[6] M. Xu, R. Niu, and P.K. Varshney, “Detection and tracking of moving objects in image sequences with varying illumination,” ICIP, vol. 4, pp. 2595-2598, Oct. 2004.

[7] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis”, IEEE Transactions on PAMI, vol. 20, Issue 11, pp. 1254-1259, Nov. 1998.

[8] F. W. M. Stentiford, “An estimator for visual attention through competitive novelty with application to image compression,” Picture Coding Symposium, pp. 101-104, Seoul, April 2001

[9] F.W.M. Stentiford, “Attention based similarity,” Pattern Recognition, 2006 (to appear).

[10] E. Batschelet, Introduction to Mathematics for Life Scientists, Springer-Verlag, pp. 157-162, New York, 1979.

[11] Multimedia Understanding through Semantics, Computation and Learning, 2005. EC 6th Framework Programme, FP6-507752, http://www.muscle-noe.org/