# MOTION DETECTION USING A MODEL OF VISUAL ATTENTION

*Shijie Zhang and Fred Stentiford*

Department of Electronic and Electrical Engineering
University College London, Adastral Park Campus, Ross Building
Martlesham Heath, Ipswich, IP5 3RE, UK
{j.zhang, f.stentiford}@adastral.ucl.ac.uk

## ABSTRACT

Motion detection and estimation are known to be important in many automated surveillance systems. It has drawn significant research interest in the field of computer vision. This paper proposes a novel approach to motion detection and estimation based on visual attention. The method uses two different thresholding techniques and comparisons are made with Black's motion estimation technique [1] based on the measure of overall derived tracking angle. The method is illustrated on various video data on and results show that the new method can extract motion information.

.

***Index Terms***— visual attention, motion analysis, object tracking

## 1. INTRODUCTION

The demand for automated motion detection and object tracking systems has promoted considerable research activity in the field of computer vision [2-6]. This paper proposes a method to detect and measure motion based upon tracking salient features using a model of visual attention.

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 and 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 has problems dealing with small and fast moving objects. It is also computational expensive. Black and Anandan [1] then proposed a framework based on robust estimation that addressed violations of both brightness constancy and spatial smoothness assumptions caused by multiple motions. It was applied to two common techniques for optical flow estimation: the area-based regression method and the gradient-based method. To cope with motions larger than a single pixel, a coarse-to-fine strategy was employed in which a pyramid of spatially filtered and sub-sampled images was constructed. Separate motions were recovered using estimated affine motions, however, the method is relatively slow. 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, but 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 because of the averaging nature of least square method.

The use of visual attention (VA) methods [7-10] 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. The visual attention approach is presented in Section 2. Results are shown in Section 3 along with some discussion. Finally, Section 4 outlines conclusions and future work.

1-4244-1437-7/07/$20.00 ©2007 IEEE

## 2. MOTION ESTIMATION BASED ON VISUAL ATTENTION

Regions of static saliency are identified using the attention method described in [9]. Those regions which are largely different to most of the other parts of the image will be salient and are likely to be in the foreground. The discrimination between foreground and background can be obtained using features such as color, shape, texture, or a combination. The concept has been extended into the time domain and is applied to frames from video sequences to detect salient motion. The approach does not require a specific segmentation process and depends only upon the detection of anomalous movements. The method estimates the shift by obtaining the distribution of displacements of corresponding salient features.

Computation is reduced by focusing on candidate regions of motion which are detected by generating the intensity difference frame from adjacent frames and applying a threshold.

$$I_{\boldsymbol{x}} = \{|r_2 - r_1| + |g_2 - g_1| + |b_2 - b_1|\}/3\,, \tag{1}$$

where parameters $(r_1, g_1, b_1)$ & $(r_2, g_2, b_2)$ represent the rgb colour values for pixel $\boldsymbol{x}$ in frames 1 and 2. The intensity $I_x$ is calculated by taking the average of the differences of rgb values between the two frames.

The candidate regions $R_1$ in frame 1 are then identified where $I_x > T$. $T$ is a threshold determined by an analysis of the image.

Let a pixel $\boldsymbol{x}$ = $(x, y)$ in $R_t$ correspond to colour components $\boldsymbol{a}$ = $(r, g, b)$. Let $\boldsymbol{F(x)}$ = $\boldsymbol{a}$. 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 $\varepsilon$ where

$$\{\boldsymbol{x}'_i \in G \quad iff \quad |x_0 - x'| \le \varepsilon\}. \tag{2}$$

Select a set of $m$ random points $S_x$ in $G$ (called a fork) where

$$S_x = \{\boldsymbol{x}'_1, \boldsymbol{x}'_2, ..., \boldsymbol{x}'_m\}. \tag{3}$$

Forks are only generated which 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 as observed in earlier work [9].

In this case the criteria is set that at least one pixel in the fork will differ with one or more of the other fork pixels by more than $\delta$ in one or more of its rgb values i.e.

$$|\boldsymbol{F_k}(\boldsymbol{x}'_i) \text{ - } \boldsymbol{F_k}(\boldsymbol{x}'_j)| > \delta_k\,, \quad for\ some\ i, j, k\,. \tag{4}$$

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 2nd fork

$$S_y = \{\boldsymbol{y}'_1, \boldsymbol{y}'_2, ..., \boldsymbol{y}'_m\} \text{ where } \boldsymbol{x_0} - \boldsymbol{x}'_i = \boldsymbol{y_0} - \boldsymbol{y}'_i \quad \text{and} \ |y_0 - \boldsymbol{x_0}| \le V\,. \tag{5}$$

$S_y$ is a translated version of $S_x$. The fork centered on $\boldsymbol{x_0}$ is said to match that at $\boldsymbol{y_0}$ ($S_x$ matches $S_y$) if all the colour components of corresponding pixels are within a threshold $\delta_k$,

$$|\boldsymbol{F_k}(\boldsymbol{x}'_i) - \boldsymbol{F_k}(\boldsymbol{y}'_i)| \le \delta_k\,, \quad k = r, g, b, \ i = 1, 2, ..., m. \tag{6}$$

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

For the *j*th of $N_I < N$ matches define the corresponding displacement between $\boldsymbol{x_0}$ and $\boldsymbol{y_0}$ as $\boldsymbol{\sigma}_j^{t+1} = (\sigma_p, \sigma_q)$ where

$$\sigma_p = |x_{0p} - y_{0p}|, \quad \sigma_q = |x_{0q} - y_{0q}|\,, \tag{7}$$

and the cumulative displacements $\Delta$ and match counts $\Gamma$ as

$$\left.\begin{aligned} \Delta(\boldsymbol{x_0}) &= \Delta(\boldsymbol{x_0}) + \boldsymbol{\sigma}_j^{t+1} \\ \Gamma(\boldsymbol{x_0}) &= \Gamma(\boldsymbol{x_0}) + 1 \end{aligned}\right\} \ j = 1, ..., N_I < N\,, \tag{8}$$

where $N_I$ 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

$$\overline{\boldsymbol{\sigma}}_{\boldsymbol{x_0}}^{t+1} = \frac{\Delta(\boldsymbol{x}_0)}{\Gamma(\boldsymbol{x}_0)}\,. \tag{9}$$

A similar calculation is carried out between $R_t$ and $R_{t-1}$ (swapping frames) 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 process 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

### 3.1 Road Scene

A pair of 352x288 frames from a traffic video was analyzed with results shown in Figure 1. The intensity difference indicates the areas of candidate motion for subsequent analysis. Motion vectors were calculated as above for each pixel in the car region and plotted in Figure 2. A map for motion magnitudes in the y direction is shown in which colors represent magnitudes as indicated in the colour bar. The directions of pixel motion is also represented by colors in the bar where angles are measured anticlockwise from the vertical e.g. yellow indicates a direction towards the top left. Motion vectors are not assigned if no internally mismatching forks can be found e.g. in areas of low saliency. The processing took 0.23 seconds in C++.

The parameters of the experiment were $M$ = 100, $N$ = 100, $\varepsilon = 3$, $m$ = 7, $V$ = 10, $\boldsymbol{\delta}$ = (40,40,40), $T$ = 12.6. $T$ is set to be twice the standard deviation of the values in the matrix $I_x$.

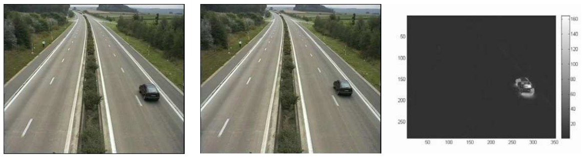

**Fig. 1.** Two adjacent frames and their intensity difference

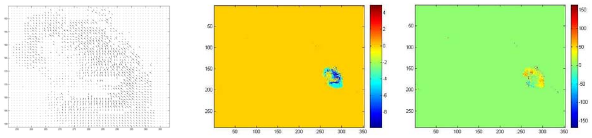

**Fig. 2.** Motion vector map, y direction magnitude map, and angle map corresponding to frames in Fig. 1.

A second pair of frames from the same traffic video was analyzed with results shown in Figure 3. The figure includes a magnitude map and an angle map for 5-car scenario.

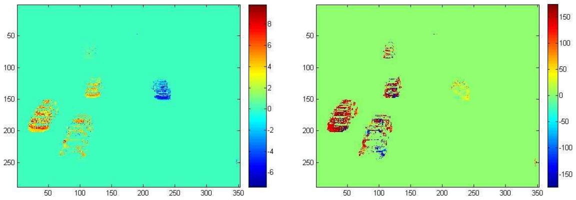

**Fig. 3.** Y direction magnitude map and angle map (5-car)

### 3.2 Object Tracking

The method was compared with Black's motion estimation technique based on the weighted object tracking angle $\theta$ defined as follows

$$\theta = \frac{\sum\left(MI^2 \times AI\right)}{\sum MI^2}, \tag{10}$$

where $MI$ is magnitude of the motion of a pixel, and $AI$ is the angle of the direction of motion of the pixel. A squared weighting was used so that motion vectors with higher values have a bigger influence on the weighted angle as they are likely to be more reliable.

Colour images are used in contrast to the grayscale images used by Black because they provide more information for the comparison of regions. A new threshold $\delta$ for pixel matching was also compared based on the joint Euclidean distance of RGB colour channels rather than treating the channels separately as in (6).

In the case of pixel mismatching, there will be at least one pixel in the fork that differs by more than $\delta$ with one or more of the other pixels in the fork according to

$$\sqrt{\sum_k \left(F_k(\boldsymbol{x}_i') - F_k(\boldsymbol{x}_j')\right)^2} > \delta, \quad \textit{for some } i, j. \tag{11}$$

The fork centered on $\boldsymbol{x_0}$ is said to match that at $\boldsymbol{y_0}$ ($S_x$ matches $S_y$) if the Euclidean distance between corresponding fork pixels are all less than $\delta$

$$\sqrt{\sum_k \left(F_k(\boldsymbol{x}_i') - F_k(\boldsymbol{y}_i')\right)^2} \le \delta, \quad \forall\, i. \tag{12}$$

Figure 4 illustrates a comparison between the weighted tracking angles derived from Black's software and those derived from the motion VA algorithm using the separate and joint metrics, . The parameters of the experiment were $M$ = 100, $N$ = 1000, $\varepsilon = 3$, $V$ = 10, $\boldsymbol{\delta}$ = (40,40,40), $T$ = 12.6, and the calculations were repeated for different numbers of fork pixels ($m$) on the same frame from 2 to 49 (fully filled 7x7 fork). The ground truth angle of the car was measured to be $36° \pm 5°$. As is shown in the figure, both angle results produced by the motion VA algorithm give closer estimates than Black ($= 24.6°$). In both cases the weighted angle increases as extra pixels are added into the fork with the separate colour channel metric performing better. Increased accuracy and precision are achieved at the expense of addition computation which increases with $m$. The improvements diminish beyond 15 fork pixels.

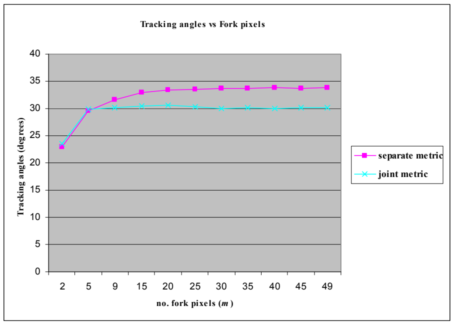

**Fig. 4.** Weighted tracking angles for Motion VA algorithm against numbers of fork pixels and separate (6) and joint (12) distance metrics. Black's estimate is $= 24.6°$.

Fork radii of $\varepsilon$ = 2,3,4,5,6 (fork sizes of 5x5 to 13x13) were then used with other parameters fixed to compare the performance on the angle estimates for motion VA algorithm using the separate channel metric. The results illustrated in Figure 5 shows that a slightly better estimate can be obtained with bigger fork radii but there is little improvement above 15 pixels. Also the results converge as the number of fork pixels increase.

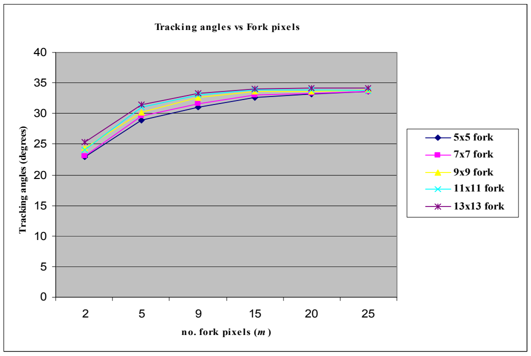

**Fig. 5.** Weighted tracking angles for Motion VA algorithm using the separate channel metric (6)

### 3.3 Football Data

The algorithm was also illustrated on football data. Figure 6 shows a pair of 352x288 frames used to estimate the motion vector map for one player in the scene. The weighted tracking angle $\theta$ was calculated to be 98.7° using separate metric as compared to the actual angle of 100°.

The parameters were $M$ = 100, $N$ = 10000, $\varepsilon = 3$, $m$ = 2, $V$ = 40, $\boldsymbol{\delta}$ = (40,40,40), $T$ = 30. The processing took 17 seconds in C++. The number of fork pixels ($m$) was set to 2 to maximize the number of matches. $V$ was increased to accommodate larger movement between the frames. A lower limit for $N$ was determined by an empirical formula which increases with both the radius of the fork $\varepsilon$ and the view radius $V$ given by

$$N \geq \left[2 \times (\varepsilon + V)\right]^2 . \qquad (13)$$

Figure 7 shows second pair of frames used to estimate the motion vector map for one player in the scene using the same parameters. It should be noted that the motion estimates include that of the camera.

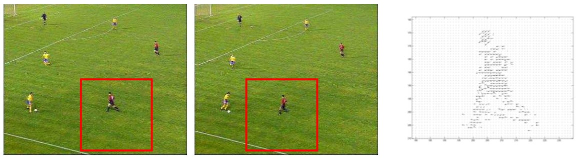

**Fig. 6.** Football frames and motion vector map

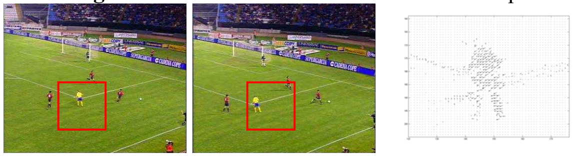

**Fig. 7.** Football frames and motion vector map

## 4. CONCLUSIONS AND FUTURE WORK

An attention based method has been proposed for motion detection and estimation. The approach extracts the object displacement between frames by comparing salient regions.

The method was illustrated on various video data and different thresholding criteria. Compared to Black's technique the attention method was shown to obtain a better estimate of motion direction. The stability of the results can be improved by increasing the volume of processing. The simple elements are amenable to parallel implementation. In addition, the method does not require a training stage or prior knowledge of the objects to be tracked.

Future work will be carried out on wider range of data to establish threshold values with more certainty with particular emphasis on addressing noise arising from background motion and changes in illumination. Camera motion involved with football data in Fig. 6 and 7 can also be estimated and used to improve motion estimation accuracy on players.

## 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] M.J. Black, P. Anandan, "The robust estimation of multiple motions: parametric and piecewise-smooth flow fields," CVIU, Vol. 63, Issue 1, pp. 75-104

[2] T. Veit, F. Cao, and P. Bouthemy, "Probabilistic parameter-free motion detection," in Proc. of CVPR, Washington, DC, USA, 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," in Proc. of ICCV, Nice, France, vol. 2, pp. 734-741, Oct. 2003.

[5] M. Kellner and T. Hanning, "Motion detection based on contour strings," in Proc. of ICIP, Singapore, 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," in Proc. of ICIP, Singapore, 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 Trans. on PAMI, vol. 20, Issue 11, pp. 1254-1259, Nov. 1998.

[8] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in Proc. of CVPR, San Diego, CA, USA, vol. 1, pp. 631-637, June 2005

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

[10] F.W.M. Stentiford, "Attention based similarity," Pattern Recognition (40), pp. 771-783, 2007

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