13—4

Selection/Substitution of Visual Features for Object Tracking

Ajith A. Pasqual, Kiyoharu Aizawa and Mitsutoshi Hatori *
Department of Information and Communication Engineering
The University of Tokyo

Abstract

In this paper we present a method of selection and substitution of visual features for carrying out object tracking. The proposed method, in principle, can make use of many visual cues available from a scene such as texture, color, velocity (monocular features) and disparity, vergence (binocular features). For the present experiments we make use of 3 features, namely, texture, optical flow and color as the main visual features and defocus of objects (implicit depth) as supportive feature. At any instance, tracking is carried out using only one feature and this feature is monitored closely for failures. The feature is substituted with another suitable feature only upon the failure or high uncertainty of the current feature.

1 Introduction

The use of multiple visual features for object tracking has been very minimal [1]. In the present paper we propose combination of three major visual features,namely, texture and optical flow, and one supportive visual feature, implicit depth/defocus (by way of blur) as a first stage of integrating visual features to track a moving target, as well as, automatic visual feature substitution. Feature substitution is done when the current feature fails to do the task.(Fig. 1)

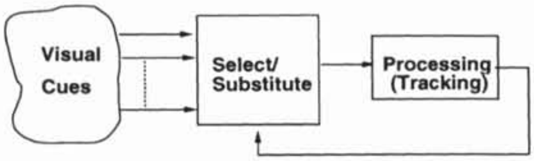


Figure 1: Feature Selection/Substitution for Tracking

The motivation for this work comes from the way human visual system tracks objects in real time and it closely matches the apparent visual substitution that human beings are capable of performing during *focused* tracking of an object. The added complexity of processing multiple features can be compensated by using simpler algorithms and software code optimizations. Increased reliability and stability offered by the use of multiple visual features makes feature substitution suitable for active vision systems and real time interactive visual tasks.

The organization of the paper is as follows: Sec. 2 describes the visual features and the way they are used for tracking. Sec. 3 gives the proposed method for visual feature substitution and Sec. 4 explains the implementational details and experimental results and finally the conclusion.

2 Selection of Cues

The human visual system makes use of a number of features of the target of interest in order to follow it. **Monocular Cues** are defined as those features of the target that could be calculated using monocular image frames, such as, motion (optical flow), texture, color and degree of blur on the image (this implicitly represent depth). **Binocular Cues** are disparity and vergence Information both of which provide an accurate measure of depth. However the use of binocular features requires stereo image sequences.

The absolute depth information from defocus is not determined. Instead this feature is used to make the logical decision as to whether the selected object in the current frame is at focus or not. This information is then used to determine the correct target in case of incorrect target selection.

2.1 Texture/Intensity

The gray scale histogram of an image is an unique way of representing the distribution of texture/intensities across an image. This information could be efficiently used for matching a target object with a candidate region in the current frame. This matching is performed using **H**istogram **I**ntersecton (HI) described in [2]. A region R can be represented by a function $f(i, n_i)$ where i is the gray scale value and n_i represents the corresponding normalized number of pixels. If regions M and R are represented by the two histogram functions $f(i, n_i)$ and $g(i, n_i)$ respectively, the Histogram Intersection Γ is defined as

[*] Address: 7–3–1, Hongo, Bunkyo-ku, Tokyo 113, Japan. Email: pasqual@hal.t.u-tokyo.ac.jp

$$\Gamma = \sum_{i=0}^{255} (Min[f(i, n_i), g(i, n_i)]) \qquad (1)$$

- Γ will always be less than 1.0.
- Region M can be considered as the target and taken as the model that needs to be matched.
- Higher the value of Γ, higher will be the similarity of the two regions.

If Region R_j is any region extracted from the neighbourhood search window for comparison with M and then region R_j corresponding the $Max[\Gamma(M, R_j)]$ is selected as the matching region. The advantage of using this method is that any block size can be used as long as it tightly encloses the target. The method will fail if the target occupies only fraction of the target window. Also the method is robust to 2D rotation as the gray scale distribution remains the same.

2.2 Color

Color is the strongest of the visual attributes of an image. The color in this work is treated as an extension of gray scale image analysis and represents an increase in dimensionality in the form of RGB color space. The increased dimensionality represents an increase in the complexity. However the inherent parallelism in RGB space offers a significant speed improvement by using software threads for color image processing.

2.3 Optical Flow

Optical flow velocities are calculated using Lucas and Kanade algorithm [3]. The main reasons for using this algorithm is its simplicity of implementation and higher computational efficiency. Also, due to its use of first derivative of the image gradients to calculate flow velocities, the extracted velocities are inherently robust to noise which makes it quite suitable for real time tracking. U and V velocity histograms are used to find the representative velocities of the search region and also similar velocity regions are marked.

2.4 Implicit depth as a visual cue

Image blur can be used to identify the correct target if there are two or more competing regions at different depths. Assuming that the two regions are at different depths d_1 and d_2 with intensity functions $I_1(x, y)$ and $I_2(x, y)$, $I_t(x, y)$ is the intensity function of the original target, the blurred region of the original target is given by

$$\hat{I}_t(x, y) = I_t(x, y) \otimes PSF \qquad (2)$$

where PSF is a Gaussian mask. Then histogram intersection is carried out between $I_1(x, y)$ with $\hat{I}_t(x, y)$ (blurred original target) and $I_2(x, y)$ with $\hat{I}_t(x, y)$ to determine the correct target. If the original target is at depth d_1 and in the extracted regions the target is slightly defocussed, matching with the blurred original target (high spatial frequencies eliminated), gives higher value for histogram intersection than for an object which is in focus.

Fig. 2 shows the process of using implicit depth to identify the target in case of a failure point. Here the histogram intersection is carried out between the extracted object in the current frame and the blurred original target in the first frame. This method is successful when the identified object and the correct target are at two different depths at the current frame. Presence of a similar object at the same depth will result in a failure.

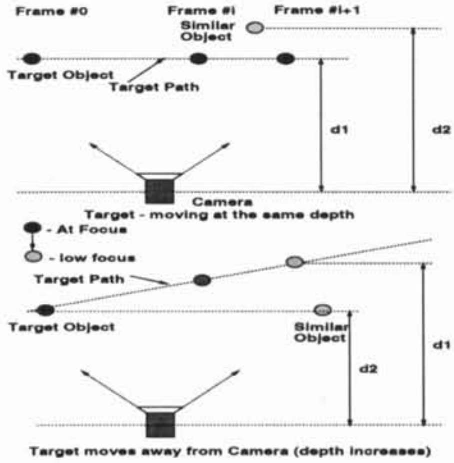


Figure 2: Using Focus Information (Implicit Depth) for recovering the tracking path

3 Selection/Substitution Process for Visual Features

Fig. 3 gives the basic flow diagram of the proposed method for visual cue selection/substitution.

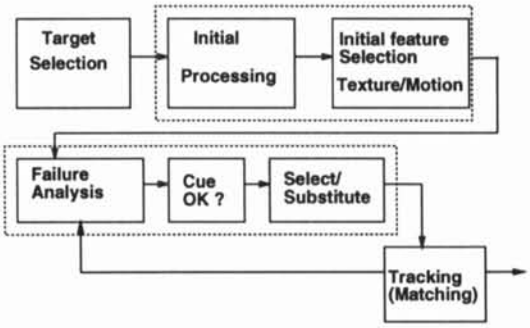
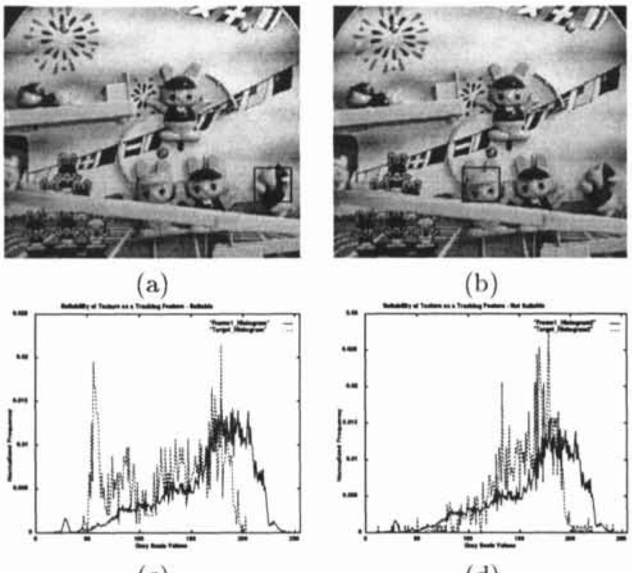


Figure 3: Basic Flow Diagram

Stage 1 is the selection of the target object by the user. At *Initial Processing* stage various parameters that can accurately represent the object are calculated (such as the histogram bin, gray scale spread, mean, histogram bin of the blurred original target etc.,). In case of color, this is the calculation of RGB histogram. Then optical flow is calculated using the first few frames and mean flow vector as well as similar velocity regions, if any are determined.

In the next stage, which can be called *Suitability Analysis*, an initial selection of feature is made using which tracking of the target is started. The gray scale distribution of the initial frame and that of target is used to automatically determine whether a technique such as histogram matching is suitable for tracking. If the histograms of the target in the first frame and the entire frame have distinctive peaks with a sufficient difference (above an empirically determined threshold), then texture is selected as a viable feature. If the histogram distributions have a near identical shape, (determined by the difference between their peaks and the variance), the texture can be eliminated as a feature. Fig. 4(a) and (b) show two targets mark in the same image and 4(c) and (d) show the corresponding histograms. Each contains the histogram of selected target and the entire image. Distinctive peaks of 4(c) and the larger difference between them suggests that texture is suitable for tracking the marked target. Similar analysis indicates that texture is not suitable for tracking the target in 4(b).



(a) (b) (c) (d)

Figure 4: (a) Texture suitable to track target (b) (c) Texture not suitable to track the marked target Gray-scale Histogram corresponding to (a) (d) Gray-scale Histogram corresponding to (b)

If the initial inter frame calculations give more than one similar velocity regions, then this stage will decide to use an Intensity Index such as Histogram Intersection Value for verifying that the correct region is extracted. Initially user intervention is necessary to select the suitable set of features. E.g. if the image sequence contains many moving objects in random directions, then tracking using optical flow may not succeed. Texture in this case offers more stability.

At any instance of time, a single feature most suitable for the scene, is used and the performance of this feature is monitored for failure points. The failure points are identified by using one or combination of 3 techniques, namely, displacement, mean squared difference and angle histogram matching [4]. At the failure point, 2 visual features texture and degree of blur are used to get back to the correct track. Refer Sec. 2.4 for more details on using implicit depth. Failure point in case of texture means that tracking algorithm has identified a similar texture object in the vicinity or in case of optical flow, another object with a similar velocity. By comparing Histogram Intersection values of the extracted regions with the original target object, target in the current frame can be identified. A special feature of this concept is the efficient use of texture, by way of **H**istogram **I**ntersecton (HI) , as a visual feature. Although the texture is quite vulnerable to change of lighting, the results are quite impressive. If a feature continuously fails, then tracking is performed using another feature. E.g. if the initial tracking is done using optical flow and it fails then texture takes over. This is done by keeping a counter for failure points.

Examples of Visual Feature Substitution

A few instances where information of one cue could correctly substitute another to continue tracking a given object are given below :

- **Optical Flow vs Intensity/Texture**
- **Optical Flow vs Focus Information**
- **Optical Flow vs Disparity** - requires a stereo image sequence

3.1 Integration of Optical Flow and Texture

The algorithm for integration of optical flow and focus information is given below. Here we make use of depth information implicitly by blurring the image. This allows us to track an object in a multiple moving object scene without using disparity which require a stereo image sequence.

- Extraction of the target (selected in the initial frame). Then this target window is blurred by convolving it with a Gaussian filter.
- Calculation of the histogram bins of the original and blurred target window. These are kept as references.
- Initial tracking of the target using optical flow.

- Optical flow will fail to track the target when there are more than one object with the same velocity in the vicinity. At this point the regions having the same velocity are extracted.

- The extracted regions are blurred using the same filter as in step 1 and HI value is calculated for each region. The region having the HI_{max} is taken as the correct region containing the target. Once the target is identified, it can be tracked again either using optical flow or HI.

4 Experiments and Results

The following section explains the experimental results obtained by applying the visual substitution algorithms and the real time implementation details.

4.1 Simulation Results

Fig.5 shows the results of algorithm applied to an outdoor scene. In this sequence, the target object (two persons walking together) is occluded twice by another person and a vehicle. Tracking successfully continues during the transition to occlusion and after the target comes out of occlusion.

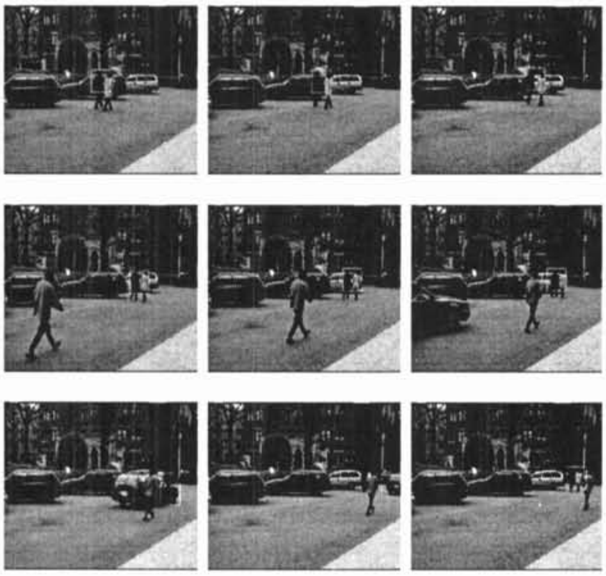


Figure 5: Tracked Sequence with 32x32 pixel Target Tracked Target in every 30th Frame is shown.

Table 1 shows a comparative statistics for 4 sequences that were used. These sequences were captured at video rate and the processing is done *off-line*. The real time implementation of the algorithms (Refer Sec. 4.2) is in progress. The number of failure points indicates the number of times tracking recovered after a feature has failed. In case of texture, the recovery is based on implicit depth and in case of optical flow it is due to texture. The low frame rate achieved for using optical flow is due to the non-optimized code used.

4.2 Real Time Implementation

A secondary objective of this work is to achieve real time tracking by using only software. In order to achieve this objective, we make use of the following:

Sequence	Avg. Frame Rate	Failure Points (Texture)	Failure Points (Opt. Flow)
Moving Doll	25	7	-
People	25	4	-
Hand Move	8	0	3
Table Tennis	25	3	-

Table 1: Sequence Statistics

- Threads - to exploit parallelism in algorithms where ever possible
- Use of Pentium MMX instructions to speed up the optical flow calculations which is the main obstacle to achieving real time operation due to its computational complexity.

Hardware Platform for Real Time Setup
BiSight Stereo camera [5] which is capable of pan, tilt speeds better than that of saccadic eye movements is used to capture images and Dual Pentium II 333MHz PC is used for processing the images as well as to control the cameras through BiSight controller.

5 Conclusion

In this paper we propose a simple but effective way of selecting and substituting visual features such as texture, color, velocity for object tracking. Simpler algorithms are used for tracking to compensate the increase in complexity of handling multiple features. Experimental results demonstrate the validity of the method. The main feature of the method is that at any point only a single feature is used for tracking and feature switching is done only when the current feature fails.

References

[1] R. Okada, Y. Shirai, and J. Miura. Object tracking based on optical flow and depth. In *Proceedings of the IEEE/SICE/RSJ Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems*, pages 565–571, 1996.

[2] M.J. Swain and D.H. Ballard. Indexing via color histograms. In *Image Understanding Workshop*, pages 623–630., 1990.

[3] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. of 7th Int. Joint Conf. on Art. Intell*, pages 674–679, 1981.

[4] A. Pasqual, K. Aizawa, and M. Hatori. Visual feature integration for real time image processing. In *Meeting on Image Recognition and Understanding (MIRU)*, pages 261–266, 1998.

[5] Carl F. Weiman and M. Vincze. A generic motion platform for active vision. In *Proc. of SPIE - Intelligent Robots and Computer Vision*, pages 435–446, 1996.