CAMERA SELECTION BASED ON ROBUST INVARIANT FEATURES FOR A TEST BED OF HUMAN-ROBOT COLLABORATION

CARLOS H. Q. FORSTER[1,2], PEDRO SANTANA[1], BRIAN C. WILLIAMS[1].

1. Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Laboratory, MERS
32 Vassar St. Room 32-221, Cambridge, MA 02139MERS CSAIL
32 Vassar St. Bulding 32-221.
Cambridge, MA 02139
E-mails: {forster,psantana,williams}@mit.edu

2. Instituto Tecnológico de Aeronáutica
Pç. Mal. Eduardo Gomes, 50. Vl. das Acácias
12228-900 - São José dos Campos, SP, Brazil
E-mail: forster@ita.br

Abstract— Cameras and artificial vision are essential elements of future manufacturing environments, where humans and robots collaborate. As cameras become abundant, there is a need to select a subset of the cameras to monitor, in order to preserve bandwidth and processing power. In this paper, we propose metrics for automatic camera selection, in support of tracking and identification of objects, which are based on robust invariant features. These metrics are analyzed using real video from simultaneous footage of a human-robot collaboration test bed.

Keywords— Computer vision, Autonomous systems, Intelligent distributed systems, Human-robot collaboration.

Resumo— Câmeras e visão artificial são elementos essenciais dos futuros ambientes de manufatura em que humanos e robôs colaboram. Frente a abundância de câmeras, há necessidade de selecionar um subconjunto delas para preservar banda de comunicação e poder de processamento. Neste artigo, propomos métricas para seleção automática de câmeras para rastreio e identificação de objetos utilizando feições invariantes e robustas. Essas métricas são analisadas utilizando video real de filmagens simultâneas de um ambiente de teste de colaboração humano-robô.

Palavras-chave— Visão computacional, Sistemas autônomos, Sistemas distribuídos inteligentes, Colaboração robô-humano.

1 Introduction

Networks of distributed cameras are important resources for monitoring the state of environments in which activities are performed involving humans and robots. Redundancy in sensing is necessary to compensate for camera limitations, such as view frustum, resolution and object visibility. As cameras become abundant, information redundancy should be managed by selecting a subset that is most relevant for the current configuration, in order to save communication bandwidth and processing power.

Although the field of Active Vision is mostly concerned with moving cameras, one advantage of switching between cameras when compared to motor-driven cameras is that the viewpoint can be changed instantaneously. While pan-and-tilt cameras cannot change their viewpoint, and thus are not useful against object occlusion, linear or rotational actuators that can change their viewpoint may move the camera too slowly, considering the activity of interest.

Camera selection is also useful when considering Data Fusion. When too many cameras are involved in the sensing, fusing all video streams is unmanageable. Additionally, robust data fusion should exclude outliers and redundant information that overloads the processing unit.

In the context of human-robot collaboration for manufacturing, it is important to have precise estimates of the state of the environment. For example, it is necessary to verify if a put-that-there task was successfully completed by a robotic agent or to interrupt an action, due to changes in the environment caused by an external agent or human counterpart.

We consider two related forms of visual information extraction that can be used for this purpose: object tracking and identification. While object tracking consists of matching points between subsequent frames of a video, object identification matches points of a template image to those of a captured video frame. Recent research in Computer Vision points to the direction of robust invariant features as a convenient means to match image points, which can then be used, for example, for tracking, identification, pose estimation, and 3D reconstruction.

Thus, a sensing architecture based on a network of smart cameras, which consist in cameras linked to a processor, can distribute the tasks of locating

robust features and computing their descriptors to local nodes and fuse the feature data in a central node.

We propose in this paper to use information from robust invariant features, in order to select the most informative camera. This will allow a system to maximize the information used for object tracking and identification, when the number of active video sources is limited.

For this, we define metrics of information quality from the robust invariant features of video frames. These metrics evaluate feature stability along the image sequence and consistency with a given template. An experiment is performed with 4 simultaneous videos of action in a manufacture test bed. Both activities of object identification and tracking are considered.

For reducing the scope in this initial effort, we consider the selection of only one best view from the 4 available. However, the results are obtained without the need of previous camera calibration, background subtraction or knowledge of any geometric model of the environment (except for the template image used for object identification).

First, we review the camera selection problem and robust invariant features (Section 2). We then detail the methods (Section 3), and present the results of the analysis of real video (Section 4). These results are discussed in Section 5 along with future directions of research.

2 Review

2.1 Camera selection

The problem of camera selection or viewpoint selection has been studied from the standpoints of both image analysis and synthesis. Many approaches, in particular those that are synthesis-oriented, assume that the selected video frames will be watched by humans and, therefore, are concerned with maintaining a stable viewpoint, while conveying as much information as possible. Analysis-oriented approaches may consider only the data gathering aspects of the problem and focus on information maximization.

Problems that should be addressed by camera selection include loss of data due to geometric considerations, such as view frustum, resolution, visibility (occlusion) and projection, as well as signal-related problems, including image noise, illumination changes, distortion and clutter. A general method for camera selection has thus proven elusive, as current methods are developed for the extraction of specific visual information.

Among the approaches to camera selection are methods centered on theory that search to optimize the portion of 3D surface represented on the images from the cameras, and coverage of the surface of objects. The probability over a visibility matrix relating camera node and visible objects is usually employed in the models and a covering subset of cameras is sought. This often results in a NP-hard problem that is approximately solved through heuristics or with imposed assumptions for simplification.

The unrealistic assumptions on knowledge of the model and requirements of camera calibration and background subtraction might make some of these approaches impractical to many real circumstances.

We list some works that represent the research in this area. Vázquez *et al.* (2001) define viewpoint entropy based on the projected area of polyhedral faces for synthetic views. Muhler *et al.* (2007) optimize the visible surface area of unoccluded surface weighted by the importance of the object and maintaining viewpoint stability of the synthetic view. Park *et al.* (2006) are concerned with viewing frustrum in large camera networks. Shen *et al.* (2007) consider the viewing angle and distance to objects, requiring calibration of cameras and background subtraction. Deinzer *et al.* (2003) use reinforcement learning with entropy as reward for tracking of objects using particle filters and the Condensation algorithm. They attempt to minimize the number of views to reach correct classification. Ercan *et al.* (2006) propose a geometric approach considering occlusion and a heuristic solution to the resulting NP-hard problem. Gupta *et al.* (2007) reduce the number of dependency relations in a Bayesian network in order to make the probabilistic approach to camera selection under dynamic occlusion tractable. The M2Tracker approach is used along with the consideration of view frustrum and dynamic occlusion. Mavrinac *et al.* (2012) employ a coverage model for view selection as a precursor to situational awareness in partially-controlled environments. Their approach is task-oriented and assumes previous knowledge of the environment.

Li *et al.* (2010) review and perform a limited comparison of techniques for camera selection. The considered tracking algorithms include particle, Kalman and Bayesian filters, CamShift, M2Tracker and histogram-based methods. For view decision, a game-theoretic technique is compared to constraint satisfaction, a fuzzy-set method and metrics of co-occurrence.

Some methods are centered on a posteriori information of the environment. Such methods may assume the presence of smart cameras that can do image processing locally, without the need to collect image data at a central point. Thus, only extracted preprocessed information is delivered to a node that will select the sources of raw image data.

Tessens *et al.* (2008) handle face detection and occlusion using an occupancy map obtained from silhouette. The employed geometric method selects the best view considering visibility of object, moving direction of object, distance from the camera, object speed and the output of the face detector. The automatic selection is compared to the selection made by a human observer.

Kelly *et al.* (2009) apply camera selection to produce a consolidated video from simultaneous footage of Tennis so specialists can analyze the details of the athlete's motions. An overhead player tracker, along with motion and dominant object detection, is used to select the best view.

Some techniques are proposed to deal with additional problems of network sensing, such as battery and bandwidth limitations. Soro and Heinzelman (2007) consider battery limitation while preserving 3D coverage in camera networks. Yang and Nahrstedt (2005) consider camera coverage of an area subject to bandwidth requirements, reducible to the NP-hard Knapsack Problem.

2.2 Robust invariant features

Matching image points is a difficult problem in Computer Vision. Salient features, such as corners and edges, have been successfully used in the past, but are not always present in images. Textural features are difficult to match under illumination changes, noise and geometrical transformations. Features proposed more recently in the literature have successfully provided invariance to translation, scale and sometimes rotation and affine transformations. They also provide some robustness to illumination changes and provide descriptors that make matching much easier. Two popular methods that provide good invariance properties are SIFT (Lowe, 1999), with emphasis on feature robustness, and SURF (Bay et al., 2006), with emphasis on both robustness and speed.

Beyond locating good features to match, these methods provide a descriptor for each feature. Descriptors are often vectors of real numbers and can be compared through the usual distances. Very close descriptors have great chance of being related to matching features. Once features are detected in both a template and a captured image (or subsequent video frames), they can be indexed by their descriptors and a fast nearest-neighbor search can be applied, in order to discover the most likely pairs of corresponding features.

3 Methods

To implement camera selection, we propose metrics based on the detection and matching of robust invariant features. For the tracking of objects, assuming there is no known model of objects of interest, the number of features matched between frames is considered. However, many of the features belong to background clutter. We, therefore, assume that the background is static and that objects of interest are most of the time in motion, particularly when they are part of the relevant action.

From our tests with SIFT and SURF features we realize that, although they have strong properties regarding invariance to geometric transformations and robustness to noise and lighting changes, many effects of the environment and the acquisition process may affect the stability of the features. While some feature points, particularly of static objects, are stable and traceable for long times, some only are detected and can be matched for a few frames and some generate spurious matching with unrelated points.

We then propose to compute the lifetime of each tracked feature to represent the stability of that feature along the image sequence. While long-living features that are static must probably belong to background clutter, those with moderate speed are usually associated to objects of interest in motion. On the other hand, recent features with high speed are probably the result of spurious matching and recent slower features are possibly the result of image noise.

To compute the lifetime of the features, we propagate the labels of the features identified in the previous image to the matching features of the current image. Unmatched features are not considered. Once a feature matches an unmatched feature of the previous image, a new label is created and the feature age is reset.

For the task of object identification, we use a template image of the object of interest to obtain its features. It is usual to filter inconsistent sets of features with robust transformation-estimating algorithms such as RANSAC, but it was not necessary in our experiment.

We consider a feature stable if it has at least a minimum lifetime and a minimum speed in the case of tracking, or if it matches a template feature in the case of identification. In both cases, the score of a view is computed as the number of stable features.

We recorded simultaneous image sequences of the same scene from four cameras. The scene contained objects of a simulated manufacturing environment, where robotic arms and humans can collaborate. The action was intended to test the limits of the feature detectors and situations in which switching the camera is needed. Additional images of objects of interest were acquired, in order to test camera selection for object identification.

The video files were processed using OpenCV implementation of SIFT features and the K-Means tree of the FLANN feature matching package (Muja and Lowe, 2009). Features were filtered considering the distance between SIFT descriptors, which are 128-value vectors. Although the selected methods are able to execute in near real time in a Core i5 computer, offline processing of recorded video was adopted for replication of the experiments with different choices of parameters.

4 Results

Figure 1 shows one original image frame and some detected SIFT features selected manually to represent objects of interest. The scene contains two robotic arms, background clutter, colored bricks with

high-contrast fiducial marks and a foam ball. As the robot arms remain static in the videos, they are considered as clutter. The identification of the object of interest (green box) is shown in figure 2. The high contrast around the fiducial marks are naturally selected by the SIFT algorithm. The foam ball was hard to track, most probably due to self shadows.

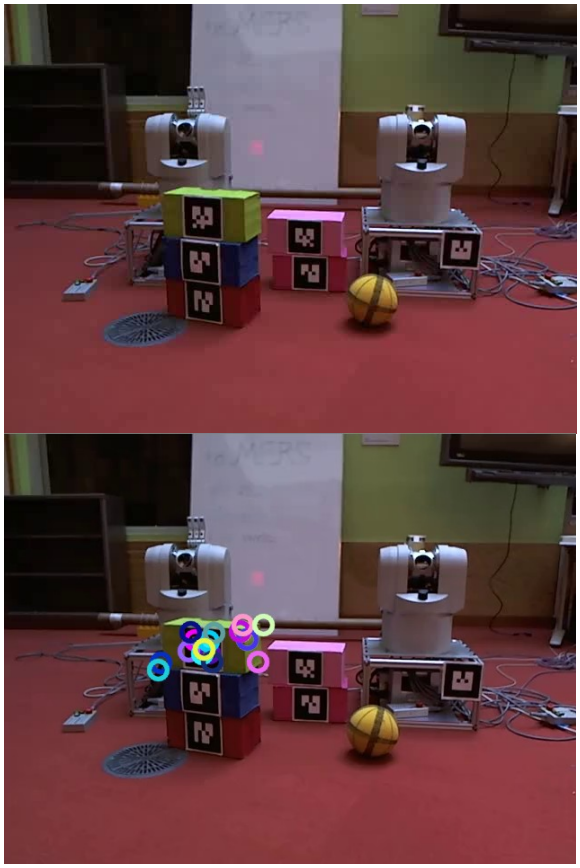

Figure 1. Initial scene and selected features to represent the object of interest: green box.

We visualize the results of tracking by plotting the detected features over the video images. Only features that have matches between frames are displayed. Also, feature lifetime is depicted as the radius of the circle representing a feature and feature speed over the image is represented as the thickness of the circle border. The color is used to identify each feature and to follow them visually. Some frames of the visualization are shown in Figure 3.

The videos captured from the four sources were synchronized and divided into 50 segments of 2 seconds. The change of camera only takes place at the end of a block in order to allow the produced video to be examined by a human. The decision to change the camera is based on the mean number of features for each frame of the segment that have lifetime greater than 2 frames and speed greater than 2 pixels per frame. The graphic is shown in Figure 4, where each line represents a different video source.

Finally, we generated a video with each segment from the best video source. Comparing the video selection from feature tracking with a human-made selection, the selected views of 18 segments are correctly pointed as first best and 14 as second best. For the selection based on object identification, one can see in figure 5, that the video does not fail to display the green box. The frames resulting of the selection based on feature tracking are shown in Figure 6.

Figure 2. Identification of the green box.

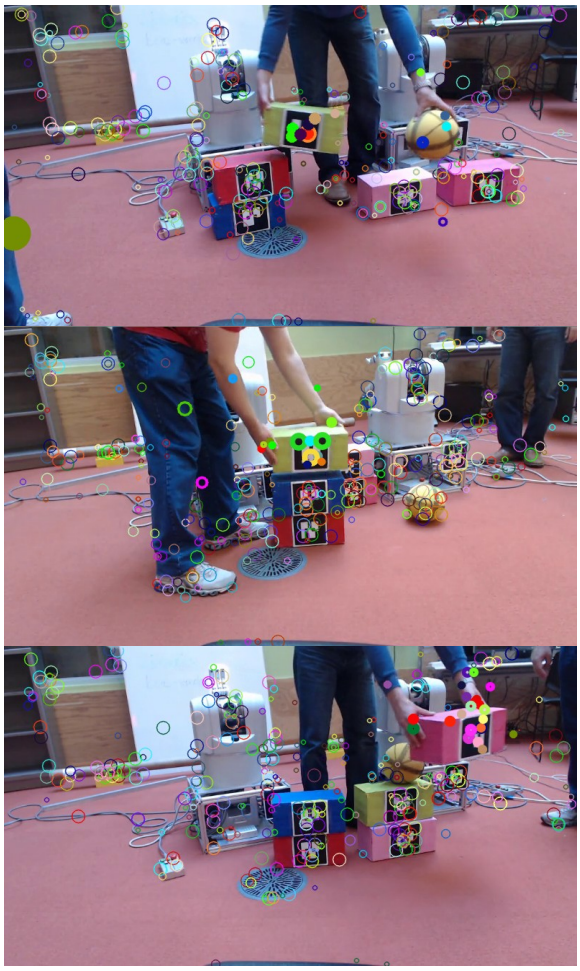

Figure 3. Visualization of feature tracking.

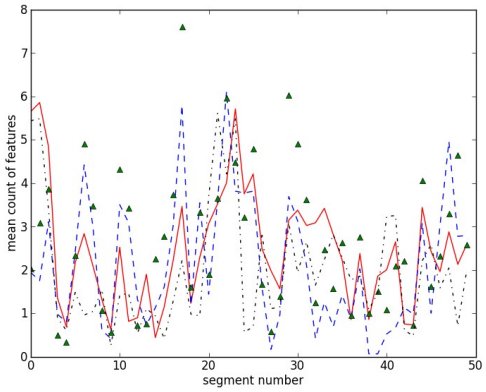

Figure 4. Selection of camera for feature tracking.

5 Conclusion

We proposed an approach to the selection of a principal view from a set of cameras based on a sensor information of the scene instead of geometric models. Previous preparation of the scene was not required at all, including geometric models, camera calibration or background subtraction. For this, information of extracted robust invariant features was used for the computation of metrics that model the importance of a given view. The model evaluates the set of detected features either on their stability along an image sequence or its consistency with an object template. The technique allows camera selection in near real time and may save bandwidth when transmitting to a central node data to allow camera selection.

Our approach is task oriented, with focus on motion tracking in video and matching to a template. For this task, the analysis of robust invariant features allowed the recognition of clutter in the scene, intervals in which objects were occluded or out of the field of view and proper detection of the areas of interest.

A future problem to address is the selection of a second best view. The difficulty of this problem is the redundant information between cameras. Once a first camera is selected, the second best camera must be evaluated on the basis of how much it improves the knowledge of the environment given that we already have the information of the first camera. So the amount of redundant information between views must be considered.

As drawbacks of our approach we may mention the disregard of additional available information and the concentration of features around areas of higher contrast. By considering the working environment to be more consistent, structured and controlled, a priori information from the scene and its devices and information from sensor fusion, which were not considered in this paper, may be used to improve the results. Evaluation with different robust invariant features might also be considered in future work.

Acknowledgments

We thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for funding this research under process number 2012/19898-4.

References

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. In Computer Vision–ECCV 2006 (pp. 404-417). Springer Berlin Heidelberg.

Deinzer, F., Denzler, J., & Niemann, H. (2003, January). Viewpoint Selection–Planning Optimal Sequences of Views for Object Recognition. In Computer analysis of images and patterns (pp. 65-73). Springer Berlin Heidelberg.

Ercan, A. O., Yang, D. B., El Gamal, A., & Guibas, L. J. (2006). Optimal placement and selection of camera network nodes for target localization. In Distributed computing in sensor systems (pp. 389-404). Springer Berlin Heidelberg.

Gupta, A., Mittal, A., & Davis, L. S. (2007, October). Cost: An approach for camera selection and multi-object inference ordering in dynamic scenes. In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on (pp. 1-8). IEEE.

Kelly, P., Conaire, C. O., Kim, C., & O'Connor, N. E. (2009, August). Automatic camera selection for activity monitoring in a multi-camera system for tennis. In Distributed Smart Cameras, 2009. ICDSC 2009. Third ACM/IEEE International Conference on (pp. 1-8). IEEE.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In Computer vision, 1999. The proceedings of the seventh IEEE international conference on (Vol. 2, pp. 1150-1157). Ieee.

Li, Y., Bhanu, B., & Nguyen, V. (2010, August). On the performance of handoff and tracking in a camera network. In Pattern Recognition (ICPR), 2010 20th International Conference on (pp. 3645-3648). IEEE.

Mavrinac, A., Rajan, D., Tan, Y., & Chen, X. (2012, July). Task-oriented optimal view selection in a calibrated multi-camera system. In Advanced Intelligent Mechatronics (AIM), 2012 IEEE/ASME International Conference on (pp. 69-74). IEEE.

Mühler, K., Neugebauer, M., Tietjen, C., & Preim, B. (2007, May). Viewpoint selection for intervention planning. In IEEE/eurographics symposium on visualization (EuroVis) (pp. 267-274).

Muja, M., & Lowe, D. G. (2009, February). Fast approximate nearest neighbors with automatic algorithm configuration. In International Conference on Computer Vision Theory and Applications (VISSAPP'09) (pp. 331-340).

Park, J., Bhat, P. C., & Kak, A. C. (2006, October). A look-up table based approach for solving the camera selection problem in large camera networks. In Proceedings of the International Workshop on Distributed Smart Cameras (DCS'06).

Shen, C., Zhang, C., & Fels, S. (2007, September). A multi-camera surveillance system that estimates quality-of-view measurement. In Image Processing, 2007. ICIP 2007. IEEE International Conference on (Vol. 3, pp. III-193). IEEE.

Soro, S., & Heinzelman, W. (2007, September). Camera selection in visual sensor networks. In Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on (pp. 81-86). IEEE.

Tessens, L., Morbee, M., Lee, H., Philips, W., & Aghajan, H. (2008, September). Principal view determination for camera selection in distributed smart camera networks. In Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on (pp. 1-10). IEEE.

Vázquez, P. P., Feixas, M., Sbert, M., & Heidrich, W. (2001, November). Viewpoint selection using viewpoint entropy. In Proceedings of the vision modeling and visualization conference (Vol. 1010, p. 01).

Yang, Z., & Nahrstedt, K. (2005, June). A bandwidth management framework for wireless camera array. In Proceedings of the international workshop on Network and operating systems support for digital audio and video (pp. 147-152). ACM.

Figure 5. Selection of camera for object identification (the green box).

Figure 6. Selection of camera for stable-feature tracking (for any object).