Augmented Exhibitions Using Natural Features

Quan Wang, Jonathan Mooser, Suya You and Ulrich Neumann
CGIT Lab, University of Southern California
{quanwang, mooser, suyay, uneumann}@usc.edu

ABSTRACT

In this paper, we propose an augmented reality application for museum exhibitions using natural features instead of calibrated fiducials to recognize paintings and recover their pose.

The proposed system utilizes an adapted Multiple View Kernel Projection (MVKP), which combines a multiple view training stage for geometric invariance and kernel projection based on Walsh-Hadamard kernels for feature description. We demonstrate that its real-time performance and robustness to lighting and viewpoint changes make it ideal for AR applications like AR exhibition systems. After obtaining the painting's index, the system retrieves related information from a remote server and displays it as virtual content overlaid on top of the painting image. Experimental results on a real-world painting exhibition have demonstrated the effectiveness of the proposed approach.

Keywords

Augmented Reality, virtual exhibition, multiple view training, Walsh-Hadamard kernel projection, object recognition.

1. INTRODUCTION

For augmented reality systems, it is essential to establish a link from the objects in the physical world to the desired displays in the augmented world. In our case, this entails recognizing an object so that we can retrieve its associated data and recover its 3D pose so that virtual objects can be accurately rendered. Most previous systems rely on tagged IDs or markers [1, 2, 3, 4]. While maker-based methods have demonstrated excellent speed and reliability, it is often difficult, if not impossible, to display a marker alongside every exhibit in an entire museum. Moreover, markers generally do not work well in the presence of occlusion.

Another line of research uses vision-based methods to determine an object's physical location and 3D pose. While some traditional single view based recognition techniques are robust and accurate enough for AR requirement, the majority is too slow for real-time applications. In recent years, multiple view image matching approaches [15, 16] received growing interests due to their real-time performance. However, to achieve the same robustness and accuracy, a large number of training views are generally needed and thus demands powerful and expensive hardware.

This paper proposes the use of natural features generated by Multiple View Kernel Projection [5]. Using Walsh-Hadamard kernels projection [6], real-time MVKP has demonstrated both effectiveness and robustness for planar objects such art paintings using a small number of training views. Additionally, as an image matching method based on local features, it naturally handles complex conditions such object occlusion and cluttered foreground or background, both typical challenges for an art museum with a large number of visitors.

The MVKP approach first builds a feature database for each painting based on a multiple view training stage. Given one input image for each painting, MVKP generates a number of synthesized affine transformed training views, detects and selects interest points, and describes local image patches around those interest points with Walsh-Hadamard kernel projection. After the training stage, Faster Filtering Vector Approximation [7] is used to establish feature correspondences between a query image and the painting feature database. Based on the object recognition result, complementary information can be retrieved from a remote server and displayed accordingly. We also introduce several important adjustments of the original MVKP method so that it will work better for the augmented exhibition system.

The remainder of this paper is organized as follows: Section 2 briefly summarizes related works. Section 3 provides an overview of our virtual exhibition system. Section 4 describes the adapted MVKP method as well as client/server information retrieval. Section 5 is the real-world painting exhibition experiment, followed by the conclusions.

2. RELATED WORKS

Augmented Reality is a natural platform on which to build an interactive museum guide. Rather that relying solely on printed tags or prerecorded audio content to aid the visitor, an AR system can overlay text and graphics on top of an image of an exhibit and thus provide interactive, immersive annotations in real-time.

Graffe, et al., for example, designed an AR exhibit to demonstrate how a computer works [8]. Their system relies on a movable camera that the user can aim at various parts of a real computer. A nearby screen then displays the camera image annotated with relevant part names and graphical diagrams.

Schmalstieg and Wagner presented a similar system using a handheld device [1]. As the user walks from place to place, AR content provides information not only about the current exhibit, but also acts as a navigational tool for the entire museum.

Conference'04, Month 1–2, 2004, City, State, Country. Copyright 2004 ACM 1-58113-000-0/00/0004…$5.00.

Both of the above systems rely on printed markers for recognition and tracking. That means that for every object to be incorporated in the AR application, a marker needs to printed and placed in the environment in such a way that it is always clearly visible. If at any time, no marker is visible inside the camera's field of view, then no AR content can be rendered. This can lead to frustrations when a particular exhibit is fully visible but its associated marker is obscured, perhaps because another visitor is standing in the way.

Our work seeks to avoid the need for artificial markers by recognizing the target objects themselves, in this case 2D drawings and paintings. Thus, as long as an exhibit is visible to the user the application can render the associated AR content.

A number of approaches have been proposed for building natural feature based AR [9][10][11]. In this paper, we use the real time image matching technique described in [5].

Our information retrieval system is based on a simplified version of the multi-tier client/server architecture described in [4]. The user interacts with a client application that recognizes exhibits and sends their unique ID numbers to the server. The server then responds with all of the relevant data for that exhibit. Thus, even with a large number of clients, content for the entire application can be controlled from a single server.

3. SYSTEM OVERVIEW

The vision-based augmented exhibition system we proposed is composed of four major components:

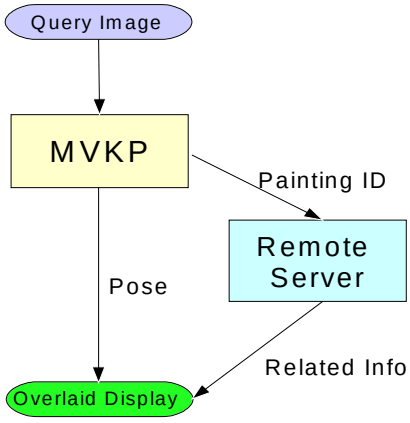


Figure 1. System overview of our augmented exhibition system

- **Acquiring query image:** the system accepts query images captured from simple camera attached to a mobile device. It can also accept single JPG images or video clips.
- **Adapted MVKP:** there are two main tasks for this component. First, given a painting, it builds the feature set for the painting. Low-resolution images (around 200x150) are enough and there is no need to extract the painting from the image in order to remove the background. Second, given the query image, it matches the painting to the database. If one of the trained paintings is matched, it establishes a feature correspondence between query image and database image. The output is the painting's ID and 3D pose with respect to the camera.
- **Remote Server:** After the server receives the painting ID through a local Internet, it retrieves the corresponding information from its database (XML file), which it sends back to the client.
- **Overlaid Display:** The client application, upon receiving the associated annotations from the server, displays them as overlaid virtual content on top of the current camera image. The virtual contents include the name of the painting and artist as well as a URL pointing to related information on the Internet. The visitor can click on the URL, which will open a web browser and bring up even more information.

Figure 1 illustrates the overall structure of our augmented exhibition system. Two major components: adapted MVKP and remote server are described in the following section.

4. MVKP AND INFO RETRIEVAL

Based on the practical requirements of the application, we chose to use MVKP as a foundation for painting recognition and 3D pose recovery. The major advantages of MVKP are: (1) robustness to lighting changes and image noises, (2) invariance to geometric distortion, (3) ability to handle complex conditions like occlusion and cluttered background, (4) sufficient accuracy for pose recovery, (5) particularly good for rigid planar objects like art paintings, (6) real-time, reliable performance, and (7) feature distinctiveness when considering a large feature database. All of these advantages make it ideal for the application of vision-based virtual exhibition system.

We also introduce several important adaptations to the MVKP method to better accommodate the requirements of AR.

In our augmented exhibition system, the outputs of MVKP method are a painting's ID and its 3D pose. The ID is then sent to a remote server through a WiFi LAN connection to retrieve the related complementary information to be displayed as virtual content on top of the painting.

4.1 MVKP method

The MVKP method in our augmented exhibition system consists of two parts: offline training and online query matching. We use Walsh-Hadamard kernel projection to describe a local image patch around each interest point. Kernel projection using Walsh-Hadamard kernels is able to measure the similarity between two sets of image patches reliably and in real time. It cannot, however, handle geometric distortions caused by viewpoint changes. To overcome this shortcoming, we combine the Walsh-Hadamard kernel projection with a multiple view training stage. The training stage provides additional information about affine distortions, allowing the same object to be matched under different view angles.

The whole method's overview is given in figure2. The details for each component are provided as follows:

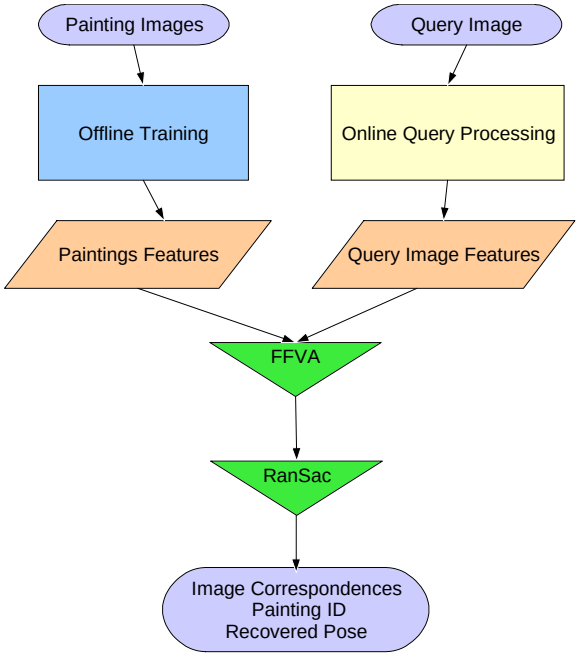


Figure 2. Structure of MVKP component in our augmented exhibition system

4.1.1 Offline training

The geometric distortion of planar object brought by 3D view-point changes can be synthesized by affine transformation. During the offline training stage, the MVKP method first synthesizes a number of training views based on the input painting image and randomly generated affine transformations. A general affine transformation can be expressed as the following [12]:

$$x' = H_A x = \begin{bmatrix} A & t \\ 0^T & 1 \end{bmatrix} x \qquad (1)$$

$$A = R(\theta)R(-\phi)DR(\phi) \text{ and } D = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \qquad (2)$$

where R is a rotation matrix and t is a translation with components t_1 and t_2. Matrix A corresponds to a rotation of θ first, followed by a rotation of $-\phi$ then scale changes of λ_1 and λ_2 in horizontal and vertical direction respectively. At last, the image is rotated back by ϕ. The six affine transformation parameters are generated randomly to cover the whole parameter space for rotation and shear angles. So we choose the ranges $\theta \in [-\pi, \pi]$, $\phi \in [-\pi/2, \pi/2]$, $\lambda_1, \lambda_2 \in [0.4, 1.6]$, $t_1, t_2 =$ 0, 1, 2, or 3.

An interest point detector will then detect potential feature points for all the generated training views. The detector searches for local maxima of the eigenvalues within 3 by 3 patches and guarantees one feature point should not be too close (for example, 3 pixels) to one another. Since we stored the affine transformations used to generate the training views, we are able to tell which feature points belonging to different training views correspond to the same physical object location. Consequently, for each physical object location we know how many times it has been detected, and thus obtain a measure of repeatability. The feature selection steps will select a number of (100-200 in our experiments) feature points with comparatively high repeatability. For each training view, we then extract 32x32 pixel image patch around each selected feature point, which we use to build descriptors for each feature. Figure 3 shows some sample extracted image patches corresponding to the same physical objection location.

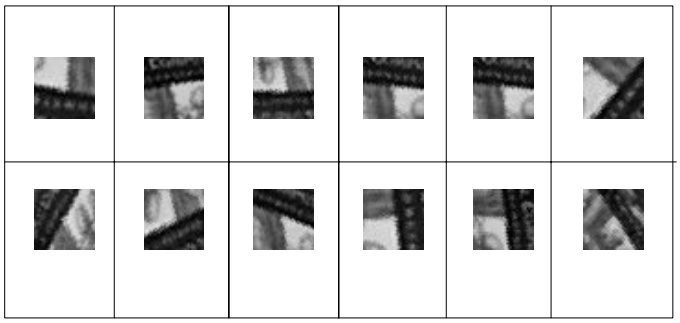


Figure3. Training image patches belonging to the same physical location

With all image patches by physical object locations, we build feature descriptors using Walsh-Hadamard kernels projection. Our strategy is to treat each patch as a vector composed of 1024 pixel intensities. These vectors can be projected onto a much lower dimensional space and still preserve the distance relationship between distinct patches.

We choose to use the Walsh-Hadamard kernels for the following reasons: (1) the Walsh-Hadamard kernels are very fast to compute and apply. One-dimensional kernels can be generated using a binary tree while consecutive kernels are α-related. Two-dimensional kernels can be generated as the outer product of one-dimensional kernels. All the coordinates of Walsh-Hadamard kernel's basis vectors are either +1 or –1. Consequently, projection onto Walsh-Hadamard kernels can be performed very fast. (2) it has been shown [13] that when the kernels are ordered according to increasing frequency of sign changes, an accurate lower bound of the original feature vectors' distance can be achieved using only a small number of kernels. Thus, we can greatly reduce the computational complexity.

Experimental results demonstrate that the Walsh-Hadamard kernels projection approach remains effective even under very noisy conditions. Furthermore, the first WH kernel corresponds to a DC value for the entire patch. By disregarding this value, we achieve a form of lighting invariance for the whole system.

The output of the offline training stage is a feature database for paintings, which consists of feature descriptors labeled by the physical object locations they correspond to. In our experiments, we typically use only the first 20 Walsh-Hadamard kernels; each feature descriptor is thus a 20-dimensional vector. K-means with 10-20 clusters is used to further reduce the size of database.

4.1.2 Online query

The online query stage first runs the same interest point detector on the query image. The only difference is, because this is an online stage designed to be as fast as possible, we only select a number of "strongest" feature points reported by the detector. After the interest point detection, the intensity values of the image patch around each interest point give us a set of feature descriptors. We normalize (translate and rescale) each patch's intensity values to the same range in order to enhance the performance against illumination changes.

We next establish correspondences between the query features and the database features. This turns out to be the most time-consuming step of online query.

Features are matched by their Euclidean distance in the 20-dinemnsional WH kernel projection space. Based on the intrinsic properties of the feature descriptors generated by MVKP, we use FFVA method [7] to do the nearest neighbor search between two feature sets. FFVA breaks down the nearest-neighbor search into two levels: 1) The *coarse search level* sequentially scans an approximations list and eliminate a large portion of data using *block distance* as lower bound, and 2) The *real data search level* calculates an accurate Euclidean distances to the remaining candidates and decides the final nearest neighbor.

Because some detected features will be located on the background or on other people or objects in the museum, clearly not all the features should be matched. Thus, *distance ratio* is used as evaluation criteria. That is: "the second closest neighbor should be significant far away from the closest one"[14]. We only accept values where the ratio of these two distances is large.

Finally, RanSac is applied to those correspondences that meet the distance ratio criteria. Because the museum paintings can be assumed to be planar, we randomly use affine transformation to fit the correspondences and keep the affinity that is consistent with the largest subset. The final outputs are a set of consistent point-based image matching and the corresponding affinity.

4.2 Adaptations

Originally, the MVKP method was used to find correspondences between two input images, which means: (1) there is no need to detect the existence of interest object and there is no search among multiple objects involved, and (2) thresholds like the one in the distance ratio criteria can be set manually since you know the query image beforehand. However, we have to make several important adaptations to the original MVKP method to meet the application requirements of augmented exhibition system.

First, for the augmented exhibition system, there can be hundreds of various painting displayed in the museum and some of them are high-textured paintings and some are not. Figure 4 shows two representative paintings. The right painting returns 50% more feature points than the left one after running the same detector. For those painting with low texture, the number of feature points returned by the detectors will also be low, which means the threshold in distance ratio criteria should also be low for it to work properly. Further more, there are other factors like feature distinctiveness of a specific painting that also affect the same threshold. And there are thresholds sharing the same dilemma other than distance ratio, for example, those thresholds in RanSac algorithm.

Figure 4. Low and high texture paintings. All artwork courtesy of Riko Conley, USC Roski School of Fine Arts

To tackle this problem, we introduce dynamic thresholds to MVKP method. Take the threshold of distance ratio criteria for example. First we set up a global goal about how many correspondences we'd like to keep after applying the distance ratio filter. At the run time, we periodically (10 times in our experiments) check the number of correspondences the method has found so far, compare it with the global goal, and adjust the threshold accordingly. Experimental results show that, with the help of the automatic adjusted thresholds, for high-textured painting we can keep the number of correspondences low and accordingly the computational cost low. For low-textured painting we will still have enough correspondences to recognize the painting and recover its pose.

Second, the user of the augmented exhibition system can point the camera to anywhere inside the museum where the query image might contain no painting at all. If there is one, we need to search and decide which painting it is. Based on our experiments, we found the size of the largest consistent correspondences set after running the RanSac is the best criteria to determine which painting, if any, is contained inside the query image.

Last but not the least, for image matching methods based on local features, especially when the query image has significant view point and lighting changes, consistent set check methods like RanSac are necessary in order to combine the global information. One problem involving RanSac in AR system is stability. RanSac randomly chooses three correspondences to fit an affine transformation, for performance consideration, terminates after a limited number of iterations. Therefore, there is no guarantee that the correct affine transform can always be found. Failure of RanSac typically means one or two frames "lose the target", which should be avoided for AR applications.

To solve this problem, we assume that when a certain painting is detected in one frame by the system, it is more likely that the same painting will appear in the following frames. In practice, after one painting is detected, the system will focus only on that painting's features in the following frames even after it encounters a RanSac failure. The system will revert to general search mode only when RanSac process fails a certain number of consecutive times. Through this implementation technique, we achieve stable and smooth displays for the augmented exhibition

system. Besides this simple technique, every frame of the input is processed independently and there are no tracking techniques involved in our current system.

4.3 Information Retrieval

Our system is based on a client/server architecture, where the client performs all of the visual processing and recognition and the server maintains a database of all known exhibits and their associated data. When the client positively identifies an exhibit, it sends a unique ID to the server. The server looks up the ID in its database and retrieves the relevant data, which may include the name of the work, the name of the artist, and possibly links to related web pages. It sends this data back to the client to be displayed over the current camera image of the exhibit.

The advantage of using a client/server model is that changes to the underlying information can be changed in one place. Whenever a client application recognizes an exhibit that it has not seen recently, it sends a new request to the server to retrieve the latest data. Due to the ready availability of wireless LAN technologies such as WiFi, it is easy to have a mobile client make periodic request to a server. Only one send-receive round trip is needed for each exhibit, so the client and server do not need to maintain a persistent open communications channel.

5. EXPERIMENTS

This section provides the experimental results for our augmented exhibition system. We use both synthesized data and real data to demonstrate the system's effectiveness. Even in the most challenging real museum test, we found that for each painting 50 generated training views and 100 features were enough for reliable performance. Other state of the art image matching methods typically require around 1000 training views [15] and several thousands features [14] for each object.

5.1 MVKP Performance

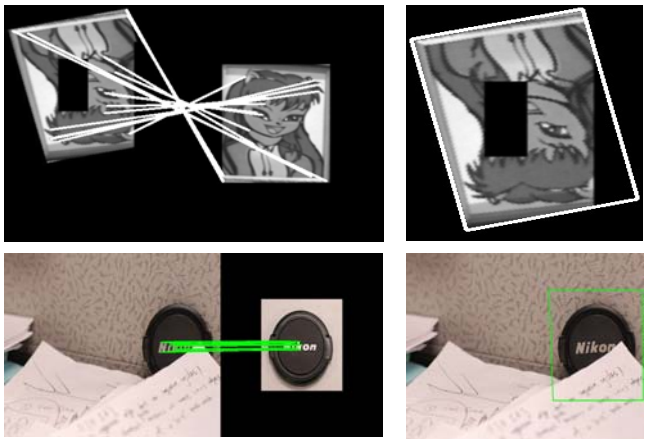


Figure 5. MVKP test, image matching and pose recovery

The core part of our augmented exhibition system is MVKP image matching. To demonstrate its effectiveness, especially compared with marker-based techniques, we tested both synthesized images with geometric distortion as well as manual occlusions and real image with cluttered backgrounds. Figure 5 illustrates some of the results. The system consistently runs at about 20 frames per second on a standard Windows PC with 3.4Ghz CPU.

5.2 Real Museum Test

Figure 6 shows the main control interface of our system. Many algorithm controls like dynamical threshold can be accessed through switches on the interface. For training the system, we first read an image from hard disk or capture a still image from the connected web camera. After clicking the training button, it takes approximately one minute to train the feature database for one object and save all the training result to hard disk. After that, when the system restarts and needs to train the same object, it will automatically load the trained feature database from hard disk, requiring only a few seconds. Our system can accept three kinds of input for query image: still image, live video from a video camera and video clip from a hard disk. The system supports five different modes to display the matched correspondences, painting recognition result and recovered pose.

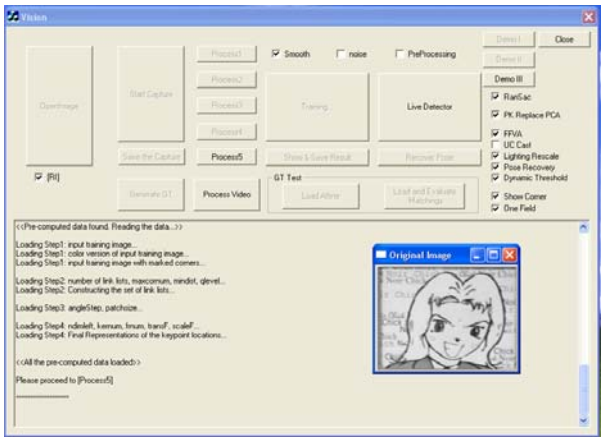
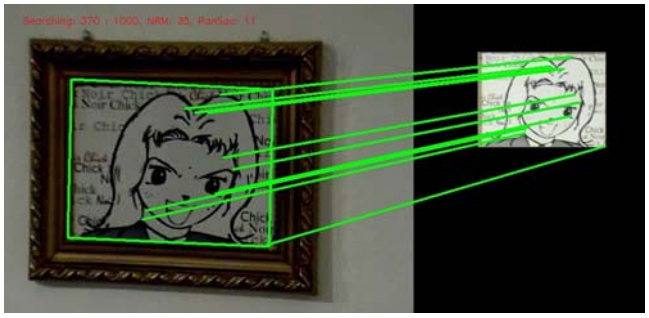


Figure 6. System main control interface

In our real museum test, we capture a video from a gallery at the USC School of Fine Art and process the video in our augmented exhibition system with four painting trained. During the video capture, we intentionally includes many challenging cases like out-of-plane rotation of the camera, moving highlights on the painting, sudden change of illumination, intense shaking of the video camera, etc. Overall, our augmented exhibition system demonstrates fast and reliable performance. Figure 7 and 8 illustrate the results with the matched correspondences displayed or hidden. A video clip showing the real-time processing together with the overlaid virtual content displayed is also available online (http://128.125.163.55/~ISMAR2007/DemoVideoQW.avi).



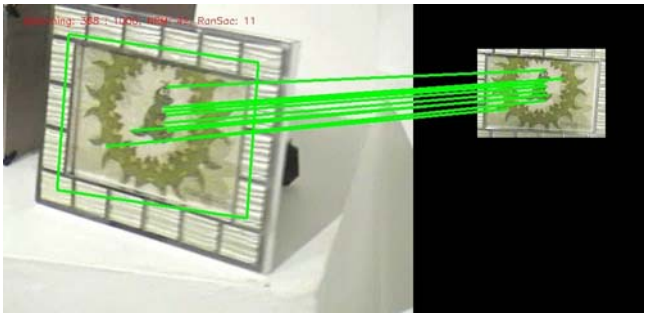


Figure 7. Real museum test, with image correspondences and recovered pose displayed

Figure 8. Real museum test, with retrieved information displayed and image correspondences hidden

6. CONCLUSION AND FUTURE WORK

This paper addresses the greatest limitation of current systems: the reliance on fiducial markers. We propose the use of natural features replacing artificial markers and demonstrate the many advantages of such an augmented exhibition system. The core part of our system is the MVKP image matching method and we introduce several important adaptations to incorporate it into the augmented exhibition system and enhance the real-world performance.

The overview of the whole system and the details of the major components are provided. Experiments demonstrate the feasibility of the proposed system under various practical and challenging conditions.

Possible directions for future work include: combining tracking techniques into the current system to obtain helpful between-frame information and adding a pre-attention stage before the MVKP to further reduce computational cost.

7. ACKNOWLEDGMENTS

This study was funded by the Center of Excellence for Research and Academic Training on Interactive Smart Oilfield Technologies (CiSoft); CiSoft is a joint University of Southern California-Chevron-initiative.
This work made use of Integrated Media Systems Center Shared Facilities supported by the National Science Foundation under Cooperative Agreement No. EEC-9529152. Any Opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.

Original artwork by Riko Conley, USC Roski School of Fine Arts

8. REFERENCES

[1] D. Schmalstieg and D.Wagner, 2005. A handheld augmented reality museum guide, in IADIS Mobile Learning 2005

[2] J. Rekimoto and Y. Ayatsuka, 2000. CyberCode: designing augmented reality environments with visual tags, in Designing Augmented Reality Environments, pp. 1–10, ACM Press.

[3] F. Ababsa and M. Mallem, 2004. Robust camera pose estimation using 2d fiducials tracking for real-time augmented reality systems, in ACM SIGGRAPH VRCAI, pp. 431–435.

[4] Mooser, Jonathan; Lu Wang, You, Suya; Neumann, Ulrich, 2007. An Augmented Reality Interface for Mobile Information Retrieval, ICME 2007 pages 2226-2229

[5] Wang, Quan; You, Suya, 2007. Real-Time Image Matching Based on Multiple View Kernel Projection, Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on , pp.1-8,

[6] Ben-Artzi, G.; Hel-Or, H.; Hel-Or, Y., 2004. Filtering with Gray-code kernels, Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol.1, pp. 556-559.

[7] Quan Wang; Suya You, 2006. Fast Similarity Search for High-Dimensional Dataset, Multimedia, 2006. ISM'06. Eighth IEEE International Symposium on, pp.799-804.

[8] Grafe, M. Wortmann, R. and Westphal, H. 2002. AR-based Interactive Exploration of a Museum Exhibit. Augmented Reality Toolkit, The First IEEE International Workshop pages 5-9.

[9] U. Neumann and S. You. 1999. Natural feature tracking for augmented reality. IEEE Transactions on Multimedia", 1(1):53–64.

[10] V. Coors, T. Huch, and U. Kretschmer. 2000. Matching buildings: "Pose estimation in an urban environment." ISAR, pages 89–92.

[11] G. Simon, A. Fitzgibbon, and A. Zisserman. 2000, "Markerless tracking usingplanar structures in the scene." In ISAR, pages 120–128.

[12] R. Hartley and A. Zisserman, 2000. Multiple View Geometry in Computer Vision, Cambridge University Press.

[13] Y. Hel-Or and H. Hel-Or. 2005. Real-time pattern recognition using projection kernels. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[14] D. Lowe. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision.

[15] V. Lepetit and P. Fua, 2006. Keypoint Recognition using Randomized Trees, Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, Nr. 9, pp. 1465 - 1479

[16] A. Boffy, Y. Tsin and Y. Genc. 2006. Real-Time Feature Matching using Adaptive and Spatially Distributed Classification Trees. In British Machine Vision Conference