應用於駕駛輔助情境之第一人稱視覺分享系統

陳少祁[*] 陳心怡[*] 陳奕麟[†] 蔡欣穆[‡] 陳炳宇[§]
[*†‡§]國立臺灣大學 [†‡§]Intel-臺大創新研究中心
[*†]{auberon, fensi, yiling}@cmlab.csie.ntu.edu.tw [‡]hsinmu@csie.ntu.edu.tw
[§]robin@ntu.edu.tw

ABSTRACT

因前方車輛所造成的視線遮擋問題是威脅行車安全的重要因素之一，解決這個問題的其中一個可能方式，是將前方車輛以第一人稱視角所看到的景象分享給後方車輛，使其視野中被前方車輛遮擋住的區域能夠經由適當的修補而還原出，去除障礙物後的景象。然而，不同車輛間的攝影機鏡頭在幾何空間上的不一致，使得車對車的視覺分享與生成變得非常具有挑戰性。在本篇論文中，我們提出了一個能夠產生第一人稱視角的影像生成演算法來解決這類的問題。首先，我們先標記出後車視野中未被遮擋的部分作為我們的參考區域，接著迭帶地從前車影像中，估計出區域單應性轉換及進行視角適應性變形，我們即可對前車影像做區域性的形變，使其視角及輪廓邊緣能夠與後車被遮擋的部份對應，並能無縫地接合在一起，讓使用者看起來似乎是前方車輛變得半透明了。我們的系統改善了駕駛者的可見度，也因此降低了駕駛過程中的負擔，進而提昇駕駛舒適度。我們以幾組在實際駕駛情境中所拍攝的具挑戰性之測試資料來展示本系統的實用性及穩定性。

Categories and Subject Descriptors

I.3.3 [**Computer Graphics**]: Picture/Image Generation;; I.4.9 [**Image Processing and Computer Vision**]: Applications;

General Terms

Algorithms

1. INTRODUCTION

Motivated by advent of cost effective and widely available camcorders, nowadays it is common to see a car driver using a dashcam (dashboard camera), a portable camera that is attached to the interior of the windshield, to record videos capturing objects in front of the car when in motion. In the unfortunate event that the car is involved in an accident, the recorded videos can serve as evidence for insurance and legal purposes. Since the dashcam can be treated as a type of first-person-view of the car, instead of using it only as a passive record, in this paper we develop a solution to utilize other vehicles' views (i.e., taken from their dashcams) to improve driver perception and increase the level of driving safety.

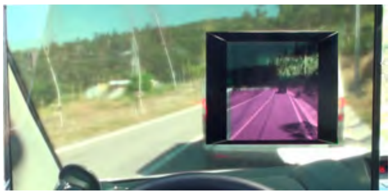

Figure 1: Previously proposed See-Through System (STS) presents to the driver a view with images taken from the preceding vehicle directly super-imposing over the image area occupied by the preceding vehicle [2]. However, the drivers need to pay extra attention since the perceived contextual information from different views are highly inconsistent.

Considering a vision-obstructing large vehicle in front of ours while driving, critical decisions such as lane changing or overtaking cannot be easily made because drivers cannot be fully aware of the potential dangers behind the visual obstruction. Although it has been shown that the overtaking vehicle can utilize direct vehicle-to-vehicle (V2V) communications to access the video data recorded by the front vehicle without significant delay [2], rendering the video streaming in the perspective of preceding vehicle requires the overtaking drivers to continuously pay attention to two disjointed views in different perspectives. Such fragmented views and inconsistent perspectives cause degradation in spatial cognition and place extra burden on the overtaking driver. For example, in Figure 1, the visual discontinuities around the boundaries of the darkened rear windscreen are not only distracting but could also have a negative impact on driving safety.

Given two synchronized video sequences I^r and I^t, which are captured by a leading vehicle (r) and the subject vehicle (t), respectively. The field-of-view in I^t are partially obstructed by the leading vehicle. Our goal is thus to generate an image sequence $\hat{I}^t$, where the occluded regions in each frame of I^t are replaced by the visible visual elements appearing in the corresponding frame of I^r with the perspective of the subject vehicle. To produce $\hat{I}^t$, a straightforward solution is to perform a pairwise image matching and stitching between two corresponding frames, as suggested by [1]. However, the performance of such process is affected by the following difficulties. Firstly, if the subject vehicle followed the leading vehicle with a short distance, the occluded areas severely downgrade the matching quality. Secondly, the inconsistent parallax from scene depth and different camera locations violate the assumptions made in typical stitching approaches [1,6]. Finally but not lastly, applying the methods designed for images to process videos may lead to temporal artefacts, e.g., the ghost effects of different misaligned objects in the video.

To address the above limitations and challenges, we propose a view-sharing system to integrate spatial information across two temporally aligned sequences. The proposed system performs both shape adjustment and color blending to generate the composited video such that the viewing perspectives and color appearances among different views are seamlessly fused and the temporal coherence can be also achieved. To this end, we propose a video-based perspective adaptation technique consisting of two main steps: *local homography estimation* and *perspective-aware warping*. With our approach, the unobstructed view and perspective of the leading vehicle can be gradually transferred and adapted to the matched occlusion region in the subject video. Specifically, our approach makes use the coherence of scene dynamics to guide the local warping across long video sequences. We also allow local homographies to be accumulated to accelerate incremental homography propagation. In addition, the parallax problem is also handled properly by restricting image stitching within a local region.

In summary, the contributions of this work are stated as follows. Firstly, we propose a view-sharing system that integrates spatial information across two temporally synchronized dashcams. The generated video sequence enables the subject driver to monitor surroundings ahead of the obstructed vehicle in accordance with current visual perception, thus providing complete situation awareness that facilitates decision making and responses to driving events. Secondly, we exploit scene dynamics in a video and propose a spatially varying warping technique for locally adapting the visibility as well as the perspective of the lead vehicle to the occluded region in the target location. It allows the subject driver to exceed the limited spatial visibility in a perspective-consistent way. Finally, we show that our system is of high practical value by evaluating it in different scenarios, including straight-lane region on highways and curved-lane region in urban areas.

2. OVERVIEW

Figure 2 depicts the overall algorithmic flow of our video perspective warping technique. The input to our system consists of a *target* and *reference* sequence, which are assumed to be temporally synchronized. The system first estimates the vision-obstruction regions (Figure 2(a)) in the target sequence (Section 3.1). To generate the corresponding contour of the visible visual region captured by the reference image, we track the robust features trajectories through the spatio-temporal volume in the reference sequence, as described in Section 3.2. The area inside the contour (Figure 2(b)) is then transferred across multiple frames by the proposed perspective adaptation algorithm (Figure 2(c)) and stitched to the matched occlusion region in the target frame (Figure 2(d)). To avoid perceptional discrepancy and mismatched boundaries between the transferred region and target image, our perspective adaptation algorithm (Section 3.3) adjusts the shape of the transferred region so that the viewpoints are continuous across the boundaries of the transferred region and the target image. Specifically, spatially-varying warping and local stitching process are performed in the area inside the contour through the video volume until the transferred region adapts to the viewpoint of the target image. Fig.2(e) shows the final synthesized image which is seamlessly composited from the reference and target images and achieves consistent visual appearance along the boundaries and perspective projection.

2.1 Problem Formulation

Considering two moving vehicles, namely the subject vehicle and the lead vehicle. Denote the *target* and *reference* sequences captured by the subject vehicle and lead vehicle as $I^t(\mathbf{x})$ and $I^r(\hat{\mathbf{x}})$, respectively. For each frame in $I^t(\mathbf{x})$, the captured scene is partially occluded by the lead vehicle. Let $\mathbf{x} = [x, y, m]$ and $\hat{\mathbf{x}} = [\hat{x}, \hat{y}, n]$ denote the spatial-temporal coordinates of I^t, I^r with $m = 1, ..., M$ and $n = 1, .., N$ indicating their frame indices, respectively. For simplicity, we will refer to the m-th target and n-th reference frame as I^t_m and I^r_n in the following discussions. Furthermore, we assume that the temporal mapping is expressed through a discrete-time signal mapping function $T : N \rightarrow R$, such that $(m, T(m))$ is an assignment of an target frame to a reference frame. For each input frame I^t_m, the temporal mapping $(m, T(m))$ is assumed to be determinable in real-time via wireless vehicular communication system. For each frame I^t_m in the target sequence, our goal is to replace the vision-obstruction region with the visual elements in the temporally corresponding frame $I^r_{T(m)}$ according to the perspective projection of I^t_m.Specifically, it will create an impression as if the lead vehicle becomes transparent, as shown in Fig. **??**.

3. METHOD

3.1 Occlusion detection

In the target sequence, the occluded areas correspondence to the lead vehicle's positions. To contour such region Ω^t_m in each frame I^t_m, the robust object tracking method proposed by Zoung *et al.* [10] is adopted to obtain an accurate vehicle position. The tracker is initialized by an vehicle object detector, then the states of the target position is estimated and updated using a collaborative model. The method outperforms many other object tracking methods [7] when the scale variation is large, i.e., the lead vehicle moves suddenly or the relative speed between two vehicles changes irregularly, which are of common situations when cars are in motion.

3.2 Contour generation

After the occluded position Ω^t_m (Figure 3(a)) in the target image I^t_m is estimated, the goal of this stage is to generate the corresponding contour $\Omega^r_{T(m)}$ in the reference image $I^r_{T(m)}$ that provides the visual elements that are invisible in the target image. We develop a forward tracking scheme in the reference video volume to find the position of such region. For each target frame I^t_m, GPS information is firstly utilized to find the spatially corresponding frame $I^r_{T(m)-k}$ (Figure 3(b)) in reference sequence, where k is an integer index offset. GPS alignment guarantees that the corresponding inter-sequence frame pair $(I^t_m, I^r_{T(m)-k})$ is taken approximately at the same geographical locations within the range about $\pm 2.5 \sim 5$ meters. Next, we perform image matching technique between the image pair $(I^t_m, I^r_{T(m)-k})$ to estimate a global transformation that related two images. Since they are captured from a similar viewpoint, a global transformation model is sufficient to roughly model their transformation. Then, we use the global transformation to locate Ω^t_m in the corresponding location $\mathbf{M}$ in the frame $I^r_{T(m)-k}$. To find the estimated contour $\Omega^t_{T(m)}$ in $I^r_{T(m)}$, starting from frame $I^r_{T(m)-k}$, we detect features within $\mathbf{M}$ and retrieves their trajectories by making a forward sweep through the reference video volume. (Figure 3(c)) shows the result of the estimated contour Ω^r. The visual elements in the estimated region $\Omega^r_{T(m)}$ is gradually transferred and aligned with the input image by the technique introduced in Section 3.3.

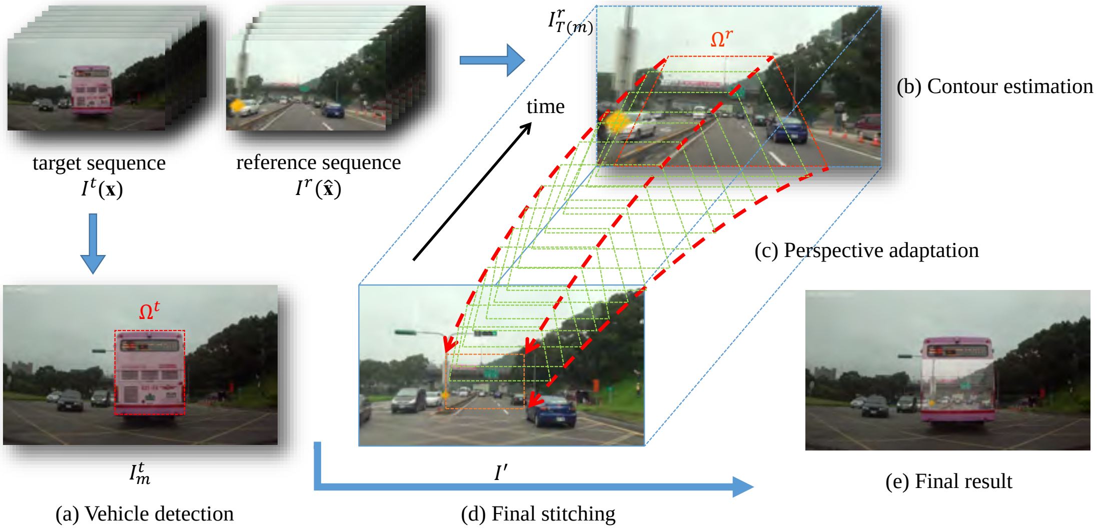

Figure 2: An overview of the proposed method. Given the target and reference sequences, the occlusion region (a) in the target image is estimated and our system automatically finds the corresponding contour (b) in the reference image. To transfer the area inside the contour in the reference image to match the occluded region in the target image, the perspective of the transferred region are adapted to fit those of the location on the target image by performing (c) perspective adaptation through reference video volume and (d) a stitching process between two image frames. In the stage of perspective adaptation, a novel view $\tilde{I}$ is synthesized by performing local homography estimation and perspective-aware warping. Finally, we stitch the synthesized view and target image where the warped region is seamlessly blended onto the target image to make an impression that the vehicle is transparent (e).

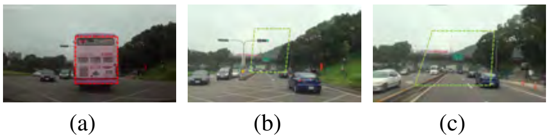

Figure 3: (a) Target image I^t_m and the occluded region Ω^t_m. (b) The spatially corresponding frame $I^r_{T(m)-k}$ of I^t_m obtained by GPS information.(c) The generated contour $\Omega^r_{T(m)}$ in $I^r_{T(m)}$ by our method.

3.3 Perspective adaptation

Given the target image I^t_m with the occlusion region Ω^t_m and the reference image $I^r_{T(m)}$ with the contour mask $\Omega^r_{T(m)}$ specifying the transferred region in the reference image to the matched occluded region in the reference image, the goal of perspective adaptation is a novel synthesized view adapting to both the shape and the perspective of the target image while closely approximating the original local appearance of the transferred region.

The perspective of adjacent pixels along the boundary of the transferred region in the synthesized image may be discontinuous if composited directly or incorporated using a simple global adjustment. Thus, for seamless transferring, a video-based perspective adaptation approach is used to adjust the viewpoint within the transferred region and remove the perspective discontinuities along the boundary while maintaining the local shape of the original transferred region.

An important characteristic of perspective projection is foreshortening: objects becomes smaller as their distance from the observer increase, as mentioned in [4]. That is, the projected size of an object depends on its depth, where the depth in a scene gradually changes when the camera is in motion. In other word, the appearance change between consecutive frames also reveal how their perspective changes. Thereby, when there is a significant discrepancy between the perspectives of the target and the reference image, the 2D shape of the transferred region must be adjusted according to the adapted motion to match such changes or discrepancies, we propose perspective-adaptation technique to accomplish this task.

3.3.1 *Perspective-aware warping and stitching*

Rendering a consistent perspective view can be achieved by estimating the transformation function between the reference and target image. Specifically, given the estimated homography $\mathbf{H} \in \mathbb{R}^{3\times3}$, a pixel at position $\hat{\mathbf{x}} = [\hat{x}, \hat{y}]^T$ in the reference image $I^r_{T(m)}$ is warped to the position $\mathbf{x} = [x, y]^T$ in the target image I^t_m by

$$\mathbf{x}' = \mathbf{H}\hat{\mathbf{x}}', \tag{1}$$

where $\mathbf{x}'$ is $\mathbf{x}$ in homogeneous coordinates. In inhomogeneous coordinates,

$$x = \frac{\mathbf{h}_1^T[\hat{x}\ \hat{y}\ 1]^T}{\mathbf{h}_3^T[\hat{x}\ \hat{y}\ 1]^T} \quad \text{and} \quad y = \frac{\mathbf{h}_2^T[\hat{x}\ \hat{y}\ 1]^T}{\mathbf{h}_3^T[\hat{x}\ \hat{y}\ 1]^T}, \tag{2}$$

where $\mathbf{h}_j^T$ is the j-th row of $\mathbf{H}$. Eq. 2 can be rewritten as:

$$\mathbf{0}_{3\times1} = \begin{bmatrix} \mathbf{0}_{1\times3} & -\hat{\mathbf{x}}'^T & y\hat{\mathbf{x}}'^T \\ \hat{\mathbf{x}}'^T & \mathbf{0}_{1\times3} & -x\hat{\mathbf{x}}'^T \\ -y\hat{\mathbf{x}}'^T & x\hat{\mathbf{x}}'^T & \mathbf{0}_{1\times3} \end{bmatrix} \mathbf{h}, \quad \mathbf{h} = \begin{bmatrix} \mathbf{h1} \\ \mathbf{h2} \\ \mathbf{h3} \end{bmatrix}. \tag{3}$$

Let $\mathbf{a}_i \in \mathbb{R}^{2\times9}$ be the first two rows of Eq. 3 computed for the i-th correspondence pair $\{\mathbf{x_i}, \hat{\mathbf{x}}_i\}$. Direct Linear Transformation (DLT) is one of the techniques to estimate the nine elements of $\mathbf{H}$ from a set of correspondences $\{\mathbf{x}_i, \hat{\mathbf{x}}_i\}_{i=1}^N$ by

$$\mathbf{h} = \arg\min_{\mathbf{h}} \sum_{i=1}^{N} \|\mathbf{a}_i\mathbf{h}\|^2 = \arg\min_{\mathbf{h}} \|\mathbf{A}\mathbf{h}\|^2, \tag{4}$$

where $\mathbf{A} \in \mathbb{R}^{2N\times 9}$ is obtained by stacking vertically $\mathbf{a}_i$ for all i. The solution is the least significant right singular vector of $\mathbf{A}$.

Although a single 2D global transformation performs well for planar scenes or rotational camera motions, but for complex scenes, i.e., highly non-planar scene that is captured by different cameras in vehicle path, as in our situation, the assumptions on motion properties and selection of dominant motions often lead to inaccurate results (Figure 4(c)). Moreover, due to the presence of occlusion, accurately aligning the input image and the reference image is much more challenging.

To tackle this obstacle, we propose a two-stage perspective-aware warping technique that utilizes the coherence of video dynamics to guide the perspective adaptation. In the first stage, we estimate the spatially varying warping functions between the consecutive frames in the reference sequence that describe how the transferred region (the area inside the contour $\Omega^r_{T(m)}$) should be deformed so that its size and shape matches the perspective of the target image. A novel view $I'_{T(m)-k}$ that integrates the visual element of transferred region while approximates the viewpoint of target image is synthesized by proceeding the process consecutively until the transferred region is gradually warped to $I^r_{T(m)-k}$, which represents the spatially closet frame of the target frame I^t_m in the reference sequence. In the second stage, we align the synthesized image $I'_{T(m)-k}$ and the target image I^t_m together, then the final composited image is recovered by blending the elements in the aligned image and the target image in the occluded region.

Feature tracking. The transformation models that relate two images are typically estimated from noisy correspondences of local invariant features. Since consecutive video frames are usually very similar, we adopt the sparse optical flow method [5] to match corresponding feature points between two neighboring frames. Sparse optical flow method estimates the motion for a selected number of pixels, thus it provides more robustness against noise than optical flow algorithms while avoids high computational cost of frame-to-frame matching by using robust feature descriptors, i.e. SIFT [3]. Specifically, we compute interest points (Shi-Tomasi features) in the video frame and generate matched points for these interest points by tracking them across multiple frames. The tracking process produces fairly accurate matching results.

Spatial varying warping function. Let $\{\mathbf{x}_i, \tilde{\mathbf{x}}_i\}_{i=1}^N$ be the collected correspondence set across consecutive frames I^r_t and I^r_{t-1} in the reference sequence, where $\mathbf{x} = [x, y]$, $\tilde{\mathbf{x}} = [\tilde{x}, \tilde{y}]$ and N is the number of correspondence pairs. To align two frames, a pixel at position $\mathbf{x}_*$ in the frame I^r_t is warped to the position $\tilde{\mathbf{x}}_*$ in the frame I^r_{t-1} by a location dependent homography model [8]:

$$\tilde{\mathbf{x}}_* = \mathbf{H}_*\mathbf{x}_*, \tag{5}$$

where $\mathbf{H}_*$ is estimated from a weighted minimization problem:

$$\mathbf{h}_* = \arg\min_{\mathbf{h}} \| \sum_{i=1}^{N} \omega^i_* \mathbf{a}_i \mathbf{h} \|^2. \tag{6}$$

subject to $\|\mathbf{h}\| = 1$ and the weights $\{\omega^i_*\}_{i=1}^N$ are calculated from a Gaussian-like distribution:

$$\omega^i_* = \exp(-\frac{\|\mathbf{x}_* - \mathbf{x_i}\|^2}{\sigma^2}), \tag{7}$$

where σ^2 is the variance. Eq. 7 gives higher weight to data points closer to $\mathbf{x}_*$. Since the problem can be written in the matrix form

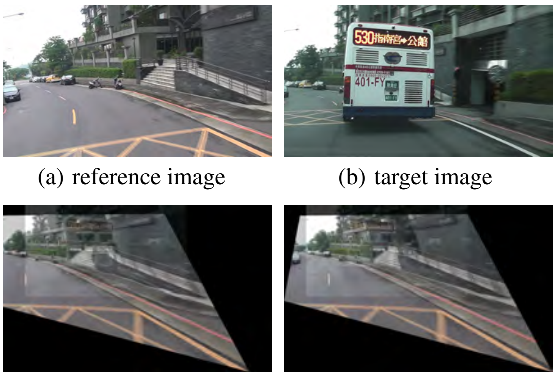

Figure 4: Aligned images. (a) reference image captured by the leading vehicle. (b) target image captured by the subject vehicle. (a) and (b) are input pairs. (c) the synthesized result stitched with global homography after final stitching. (d) the synthesized result stitched with MDLT method after final stitching.

$$\mathbf{h}_* = \arg\min_{\mathbf{h}} \|\mathbf{W}_*\mathbf{A}\mathbf{h}\|^2, \tag{8}$$

where the $\mathbf{W}_* \in \mathbb{R}^{2N\times 2N}$ can be further described as:

$$\mathbf{W}_* = diag([\omega^1_* \ \omega^1_* \ ... \ \omega^N_* \omega^N_*]). \tag{9}$$

$diag()$ constructs a diagonal matrix with a given vector. Eq. 8 corresponds to a weighted Singular Value Decomposition(WSVD) problem, and the solution is the least significant right singular vector of $\mathbf{W}_*\mathbf{A}$.

Avoiding parallax using local stitch. As mentioned in [9], the images with significant parallax often cannot be aligned well over the whole overlapping region without suffering artifacts like folding-over. To handle parallax, we also perform local stitch between I^r_t and I^r_{t-1}. Specifically, after the local homographies between frame I^r_t and I^r_{t-1} are estimated, only the area inside the transferred contour Ω^r_t is warped to I^r_{t-1}, then a novel view I'_{t-1} is composited, which correspondences to the perspective observed in I^r_{t-1}. By iteratively applying local homography estimation and perspective-aware warping between I'_t and I^r_{t-1}, the content and the perspective inside the contour $\Omega^r_{T(m)}$ is gradually adjusted. Finally, a novel frame $I'_{T(m)-k}$ is synthesized.

3.3.2 Final stitching

Owing to the first warping stage discussed in Section 3.3.1, the perspective of $I'_{T(m)-k}$ is adapted to $I^r_{T(m)-k}$, which is the spatially closest frame of the subject frame I^t_m in reference video sequence. We assume that the perspectives of two frames should be similar if the distance between their spatial coordinates is small enough. Finally, we stitch I^t_m and $I'_{T(m)-k}$ together to get the final panoramic image $\hat{I}$. In order to make the leading vehicle transparent, we only cut the part that corresponds to the occlusion mask Ω^t from the stitched view and blend it with I^t_m to get the final result $\hat{I}$.

4. EXPERIMENTAL RESULTS

We evaluate the performance of our system using video clips collected in real driving scenarios. The videos were designed to be

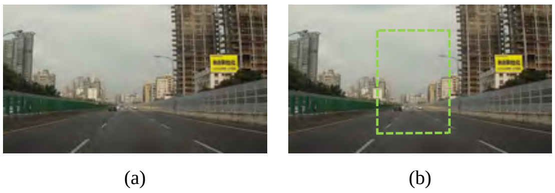

Figure 5: (a) is the frame $I^{T}_{T(m)-k}$ in the reference sequence. (In this case, we set k = 30). (b) is a synthesized frame by the proposed perspective adaptation method. One can see that the perspectives of the two different frames are very visually similar inside the mask.

captured in three different road conditions and traffic flows: (i) on the highway, (ii) on the city road and (iii) on the mountain road. The dashcams on the vehicle were set up in the middle of the windshields with timestamps information. For capturing these videos, the driver on the subject vehicle followed in the path of the bus ahead, which corresponded to the lead vehicle through our discussion. In addition, the geographical information provided by GPS receivers on the vehicles was used to spatially align two videos. Beside, we also recorded an additional video with only one dashcam. In the following sections, we first demo the effectiveness of the proposed perspective adaptation approach in 4.1. Then, we compare our method with two approaches: (i) global alignment using RANSAC and (ii) local alignment with Moving Direct Linear Transformation and demonstrate several result using the proposed method in Section 4.2.

4.1 Verification of perspective adaptation

In this experiment, we aim to use a single video sequence to validate the effectiveness of the proposed perspective adaptation approach. By gradually warping the area specified the contour captured at time $T(m)$, it perspective is matched to that of time $T(m) - k$. The result is shown in Figure 5(b). Figure 5(a) represented the ground truth perspective (the frame captured at time $T(m) - k$). It can be seen that we alter the shape of transferred region according to scene depth change to model the perspective effect.

4.2 Qualitative comparisons

We compare our method against the baseline warping method (global homography via DLT in inliers) and the local homography method (Moving-DLT) [8] with three alignment instances. Figure 6(a) and (b) show three pairs of input images with a significant amount of parallax and occlusion, where the input are the reference and target images captured by the dashcams on the lead vehicle and the subject vehicle, respectively. In each case, large viewpoint change, different lighting conditions and the presence of occlusion make the alignment task very challenging. Figure6(c)$\sim$(e) show the results generated by each method.

For the baseline method, we detect and match SIFT keypoints in the input pair, then run RANSAC to remove outliers. We estimate a global homography via Direct Linear Transformation (DLT) on inliers to align two images. The result of baseline method caused unavoidable misalignment and ghost effect, which can been seen in Figure6(c). It suggests that using a single homography alone is not sufficient to model the transformations between two images because a scene is usually composed by more than two projection planes. Besides, given input images with considerably different perspectives, it's very difficult to establish enough correct matches thus leading to incorrect homography estimation. While Moving-DLT with spatially varying homographies is able to produce good results, it tries to align two images over the whole overlapping region. Thereby, the estimated transformations are easily dominated by the noisy matches. As a consequence, the distortion is very large and ghosting still occurs in the region Figure6(d) (the stairs next to wall). In contrast, by using the coherence property in the video sequence, features are easily to be matched, it facilitate the a good in first warping stage. In the stage of second warping stage, the perspective-adapting frame which possesses similar viewpoint of the subject vehicle is transformed. Thereby, it's easier to find matches between close perspective compared with direct warping methods. Therefore, our method can estimate local homography more precisely, thus achieves more plausible results as shown in Figure6(e).

5. CONCLUSION

In this paper, we address the problem of aligning two videos that are captured simultaneously by independently moving cameras following similar trajectories. Aligning two temporal synchronized video sequences encounters great challenges due to large viewpoint changes and heavy occlusion. Therefore, in this paper we propose a two-stage warping technique that gradually adapts the perspective from one video to the other, rather than directly aligning two videos with large difference in viewpoint. It not only reduces the difficulties of perspective transferring between multiple views, but also increases the visibility of the driver and enhances safety and comfort in driving scenarios.

6. 致謝

本論文感謝科技部、國立臺灣大學與Intel Labs經費補助，計畫編號：NSC102-2911-I-002-001與NTU-ICRP-103R7501。

7. REFERENCES

[1] M. Brown and D. G. Lowe. Automatic panoramaic image stitching using invariant features. *ACM Trans. Graphics*, 74(1):59–73, 2007.

[2] P. E. R. Gomes, F. Vieira, and M. Ferreira. The see-through system: From implementation to test-drive. In *Proc. IEEE Vehicular Networking Conf.*, pages 40–47, 2012.

[3] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. Journal Computer Vision*, 60:91–110, 2004.

[4] S.-J. Luo, I.-C. Shen, B.-Y. Chen, W.-H. Cheng, and Y.-Y. Chuang. Perspective-aware warping for seamless stereoscopic image cloning. *Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2012)*, 31(6):182:1–182:8, 2012.

[5] J. Shi and C. Tomasi. Good feature to track. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[6] R. Szeliski and H.-Y. Shum. Creating full view panoramic image mosaics and environment maps. In *Proc. ACM SIGGRAPH*, pages 251–258, 1997.

[7] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2411–2418, 2013.

[8] J. Zaragoza, T.-J. Chin, M. S. Brown, and D. Suter. As-projective-as-possible image stitching with moving dlt. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2339–2346, 2013.

[9] F. Zhang and F. Liu. Parallax-tolerant image stitching. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

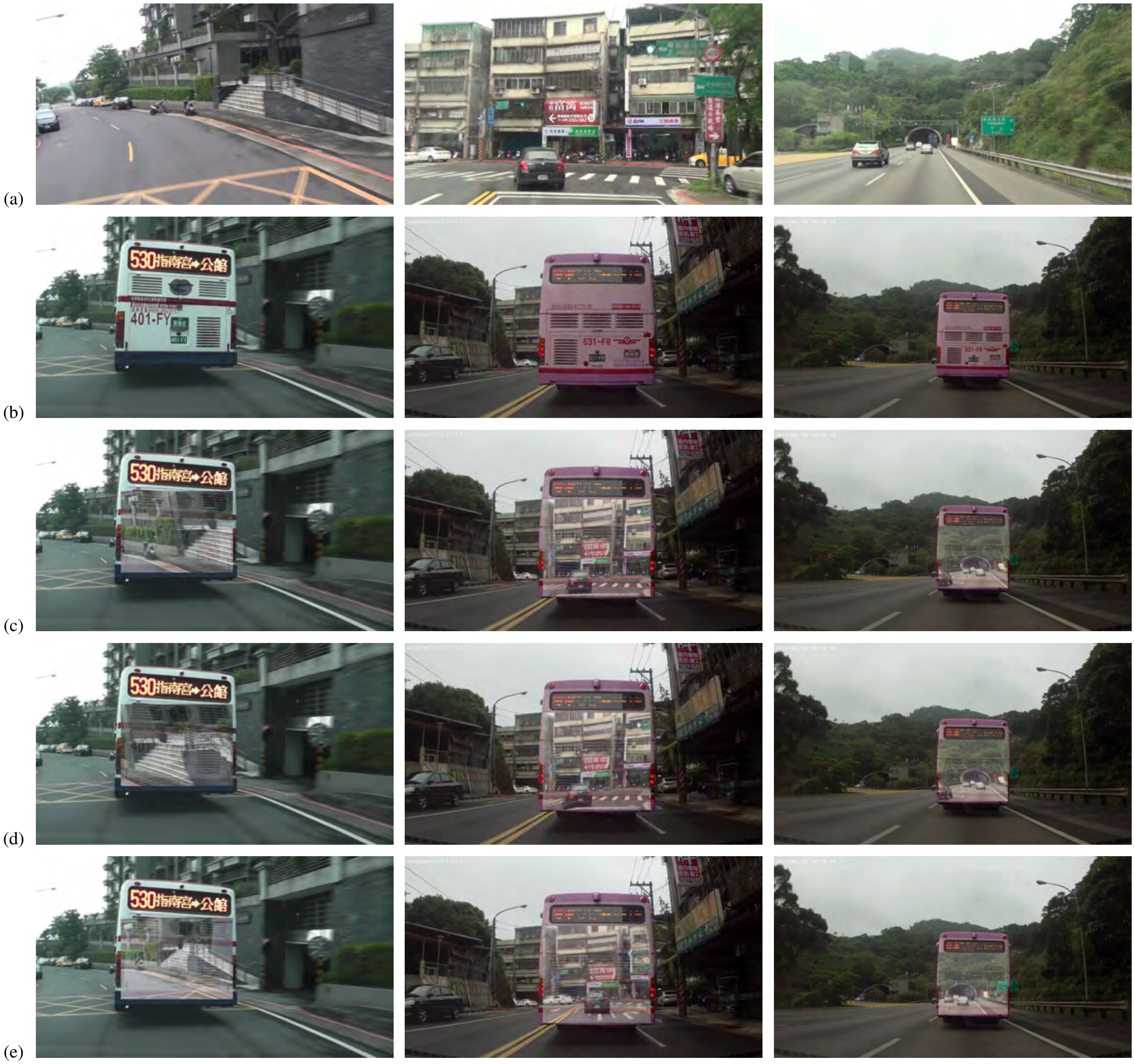

Figure 6: Qualitative comparisons. Challenging frames of reference sequence (a) and target sequence (b) are shown. (c) Aligned image using a global homography method. (d) Aligned image using Moving-DLT method [8]. (e) Aligned image with our method.

[10] W. Zhong, H. Lu, and M.-H. Yang. Robust object tracking via sparsity-based collaborative smodel. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1838–1845, 2012.