Towards Robust Night and Day Place Recognition using Visible and Thermal Imaging

Will Maddern and Stephen Vidas
School of Electrical Engineering and Computer Science
Queensland University of Technology
Brisbane, Australia
Email: w.maddern@qut.edu.au, stephen.vidas@qut.edu.au

***Abstract*—The chief challenge facing persistent robotic navigation using vision sensors is the recognition of previously visited locations under different lighting and illumination conditions. The majority of successful approaches to outdoor robot navigation use active sensors such as LIDAR, but the associated weight and power draw of these systems makes them unsuitable for widespread deployment on mobile robots. In this paper we investigate methods to combine representations for visible and long-wave infrared (LWIR) thermal images with time information to combat the time-of-day-based limitations of each sensing modality. We calculate appearance-based match likelihoods using the state-of-the-art FAB-MAP [1] algorithm to analyse loop closure detection reliability across different times of day. We present preliminary results on a dataset of 10 successive traverses of a combined urban-parkland environment, recorded in 2-hour intervals from before dawn to after dusk. Improved location recognition throughout an entire day is demonstrated using the combined system compared with methods which use visible or thermal sensing alone.**

I. Introduction

For long-term vision-based robot operation in outdoor environments, the ability to cope with the extreme illumination changes due to the day-night cycle of the sun is a crucial requirement and remains a largely unsolved problem. Current state-of-the-art systems which function in outdoor environments typically use scanning LIDARs [2] [3], appearance-based LIDAR [4], or active illumination [5], but these active systems have significant power requirements and weight penalties. Despite a significant body of work towards illumination-invariant image recognition [6] [7], passive vision sensors operating in the visible spectrum suffer from the fundamental limitation of cyclic appearance change over a 24-hour period [8] as well as gradual change over longer timescales [9].

Passive thermal-infrared imagery is not directly affected by changing lighting conditions, and has demonstrated robustness to many environmental conditions in the past [10]. However, thermal-infrared imagery suffers indirectly from cyclic illumination; the contrast between objects with different thermal properties varies as they are heated by the Sun and cooled by thermal radiation throughout the day. The phenomenon known as 'thermal crossover' [11] describes the time of day when objects in a scene exhibit minimal thermal contrast, which complicates place recognition when using thermal-infrared images alone.

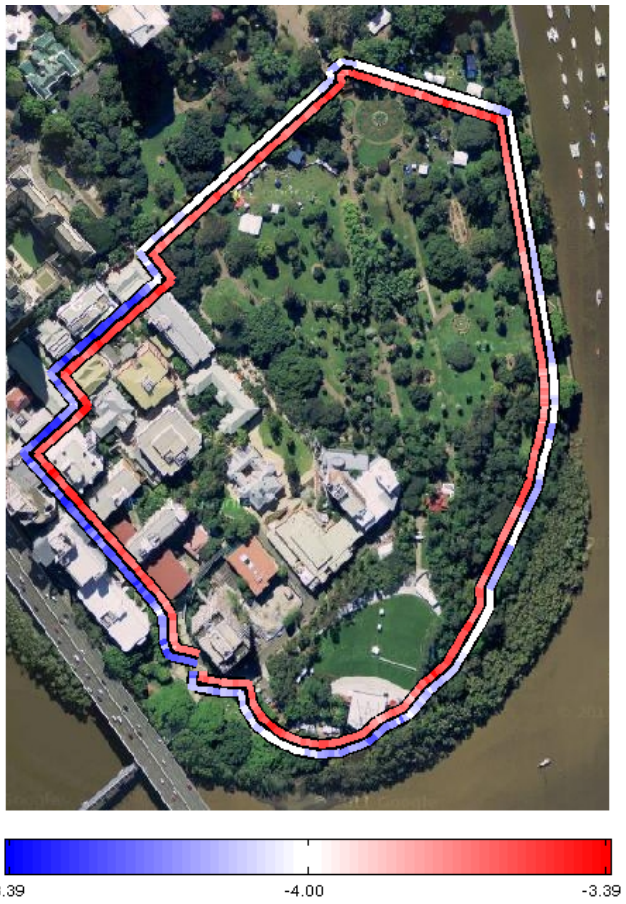


Fig. 1. Route map with observation log-likelihoods between the 5am and 7am sequences. Red regions represent locations with high log-likelihoods in the thermal-infrared modality, while blue represents those with high log-likelihoods in the visible modality (more color indicates higher likelihood). The superior stability of the thermal-infrared modality under illumination change is apparent.

In this paper we investigate a combination of visible-spectrum and thermal-infrared imagery to perform loop closure detection over the course of a full day. The complementary nature of visible-spectrum and thermal-infrared has the potential to improve the performance of existing visible- or thermal-only systems. We detect loop closures using FAB-MAP [1] and propose a method for fusing the two modali-

ties in the Chow-Liu tree representation to achieve superior location-recognition performance. The fused representation is evaluated on a multiple-modality dataset consisting of 10 successive traversals of a combined urban-parkland environment. The traversals were performed over 20 hours, spanning from before dawn to after dusk.

II. Background

A. Image processing

Current state-of-the-art feature detector and descriptor extraction algorithms provide some level of contrast and illumination invariance designed for visible-spectrum images. However, as demonstrated in [12] these methods do not naively extend to thermal-infrared images. Raw thermal-infrared images are typically stored in 14-bit per pixel format, with the number of active gray levels varying considerably depending on the scene. In many images, the span of active gray levels will be less than 255, and in these cases the image can easily be converted to 8 bits per pixel making them easier to interface with conventional computer vision algorithms. In images with more a span of more than 255, some quantization may be required. In general, the level of contrast in thermal-infrared images will be significantly less than visible images. This can result in low feature counts from local feature detection algorithms if special care is not taken to adjust detector sensitivities accordingly. Real-time adaptive contrast enhancement algorithms such as those presented in [13] and [14] provide a means for adapting thermal-infrared images to perform better with normal feature detection and description algorithms.

B. Loop Closure Detection

In this paper we use the FAB-MAP algorithm [1] to determine loop closure events based on the set of features detected in each image. Each image is converted into the visual bag-of-words representation described in [15]. It is necessary to create a database of common features from a set of training data in a similar environment to the test environment prior to performing localisation [1]. Every feature extracted from the image is converted to the closest visual 'word', reducing each image to a binary vector of which visual words are present in the image.

$$Z_k = \begin{Bmatrix} z_1 & z_2 & \dots & z_n \end{Bmatrix} \tag{1}$$

Each unique location L_k is represented by the probability that the object e_i (that creates observation z_i) is present in the scene.

$$\begin{Bmatrix} P(e_1 = 1|L_k) & P(e_2 = 1|L_k) & \dots & P(e_n = 1|L_k) \end{Bmatrix} \tag{2}$$

The probability of a new image coming from the same location as a previous image is estimated using recursive Bayes:

$$P(L_i|\mathcal{Z}^k) = \frac{P(Z_k|L_i, \mathcal{Z}^{k-1})P(L_i|\mathcal{Z}^{k-1})}{P(Z_k|\mathcal{Z}^{k-1})} \tag{3}$$

where Z^{k-1} is a collection of previous observations up to time k. The prior probability of matching a location $P(L_i|\mathcal{Z}^{k-1})$ is estimated using a naive motion model. The observation likelihood $P(Z_k|L_i, \mathcal{Z}^{k-1})$ is assumed to be independent from all past observations and is calculated using a Chow Liu approximation [16]. The Chow Liu tree is constructed once as an offline process based on training data. Observation likelihoods are approximated using the Chow Liu tree as follows:

$$P(Z_k|L_i) \approx P(z_r|L_i) \prod_{q=1}^{n} P(z_q|z_{p_q}, L_i) \tag{4}$$

where r is the root node of the Chow Liu tree and p_q is the parent of node q. The observation likelihood is the crucial component for determining the match between two individual images; it incorporates the features present in both images along with the Chow-Liu environment model to calculate the likelihood that both images represent the same location.

For FAB-MAP to correctly recognise locations there must be some overlap in visual words between images. Aside from special cases where the Chow-Liu tree captures sufficient feature co-occurrance information to match images containing no common words, the recall performance of FAB-MAP is generally enhanced by increasing the number of common visual words between images.

The remainder of this paper will address how images from the thermal-infrared modality can be used to increase visual word repeatability, and therefore improve the observation likelihood scores between images of the same location at different times of day.

III. Dataset

Data was collected using a platform consisting of a GPS receiver, visible-spectrum camera and thermal-infrared camera, as shown in Figure 2. The platform was attached to the rear tray of a bicycle so that the two cameras faced directly rearwards, parallel to the ground. GPS positioning information was unreliable due to the nature of the environment (which included tall buildings and significant tree cover) so ground truth positions were manually corrected.

The thermal-infrared camera used for the experiments was a Thermoteknix Miricle 307K, which consists of a long-wave uncooled microbolometer detector sensitive in the 7- to 14-µm range. The camera has a resolution of 640 by 480 pixels and a horizontal field of view of approximately 60 degrees and is rated for objects in the temperature range of $-20°$ to $150°$ Celsius. The camera has a noise-equivalent differential temperature (NEDT) of 85 mK. Fourteen-bit monochromatic images were captured at 15Hz over a USB connection.

A Point Grey Grasshopper2 1394b camera with Bayer filter was used to collect visible-spectrum images with identical resolution and approximately identical FOV to the thermal-infrared images. The visible-spectrum images were temporally aligned with the thermal-infrared images within route traversals based on time-stamp information, and manually aligned between route traversals.

Fig. 2. Capture platform, with (from left to right) the GPS receiver, thermal-infrared camera and visible-spectrum camera.

The route for the data capture was an approximately 1500m long stretch in and around the QUT Gardens Point campus and consisted of substantial portions of both urban environment and parkland, pictured in Figure 1. Ten traversals with two hours separation between each were undertaken, spanning from 5.00am to 11.00pm. As shown in Figure 3, this represented a period lasting from before dawn to after dusk. The full dataset of images from both modalities was subsampled to 300 frames per route traversal per camera, resulting in a total of 6000 frames across the full experiment. This corresponds to image pairs with approximately 5m spacing throughout all the route traversals. Frames were manually aligned between datasets to ensure approximately identical viewpoints.

IV. Methodology

The following section outlines the steps taken to determine localisation performance across modalities across a full day. The first step involved determining an appropriate pre-processing scheme and feature extraction method for each modality, along with extracting descriptors for each feature. The second step involved constructing a visual vocabulary for the bag-of-words representations and Chow-Liu tree to capture conditional likelihoods for feature co-occurrence, then using FAB-MAP to perform place recognition.

A. Feature extraction

Raw output from the visible camera was already in an 8-bit per pixel format appropriate for feature detection and description, and so no pre-processing was required beyond converting to grayscale for simplification. For the thermal-infrared images, when the span of raw intensity values was less than 255, conversion to 8-bit format consisted simply of shifting the intensities so that the median pixel value was 128. In images where the span of raw intensity values was greater than 255, the top and bottom 0.1% of intensity values were thresholded, and all values were then linearly mapped to the interval $[0, 255]$. The CLAHE (Contrast Limited Adaptive Histogram Equalization) [13] algorithm was then used to enhance the contrast of the thermal-infrared image, with a normalized clipping limit of 3.8.

For both modalities the OpenCV[1] implementation of the SURF [7] feature detector and descriptor was used to extract local feature descriptors from the images. The detector was implemented in its dynamically adapted mode, meaning that upper and lower bounds for feature counts could be specified. Limits of 500 and 1000 features were chosen for the visible modality, with a maximum feature radius of 80 pixels, while for the thermal modality limits of 600 and 800 were selected and a larger radius of up to 100 was permitted. These parameters were found to achieve the most promising performance in terms of word overlap in the initial development stage.

B. Place recognition

The visual vocabulary is generated using k-means resulting in a 5,000 word codebook for each modality from the 1pm dataset. We then match each descriptor in every image to the nearest cluster centre in the vocabulary; for each location, we obtain a pair of visual bag-of-words representations as follows:

$$Z_v = \begin{bmatrix} z_{v_0} & z_{v_1} & \dots & z_{v_n} \end{bmatrix}; \; Z_{th} = \begin{bmatrix} z_{th_0} & z_{th_1} & \dots & z_{th_n} \end{bmatrix} \quad (5)$$

A Chow-Liu tree [16] was constructed for each of the two codebooks, to capture the conditional dependencies between words independently for each modality. To compare locations we use the FAB-MAP 2.0 implementation in the openFABMAP[2] repository, since the original FAB-MAP binaries do not allow the use of custom codebook or Chow-Liu tree representations. Additionally, using openFABMAP we can extract the observation likelihoods of equation 4 directly without requiring the full recursive Bayes estimation of equation 3.

C. Combining representations

In order to combine the sensor modalities at the bag-of-visual-words level (rather than attempting fusion of the images themselves), we concatenate the individual word vectors for each image as follows:

$$Z_c = \begin{bmatrix} z_{v_0} & z_{v_1} & \dots & z_{v_n} & z_{th_0} & z_{th_1} & \dots & z_{th_n} \end{bmatrix} \quad (6)$$

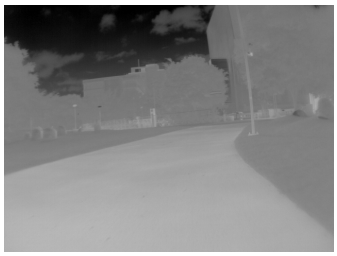
Using this combined bag-of-words representation, we train a Chow-Liu tree between modalities. Crucially, this allows us to not only take advantage of modelling feature uniqueness within modalities, but also to capture the distribution of conditional observation likelihoods between visible-spectrum and thermal-infrared images, allowing us to infer location matches from the combined visible-thermal representation even in the absence of direct feature matches from either modality. The concatenated bag-of-words representation consists of 10,000 visual words.

[1] http://opencv.willowgarage.com/wiki/
[2] http://code.google.com/p/openfabmap/

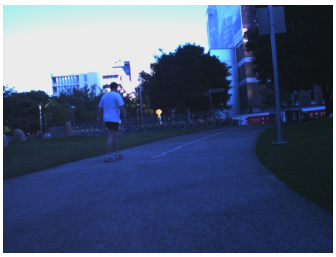
(a) 5.00am - visible

(b) 5.00am - thermal

(k) 3.00pm - visible



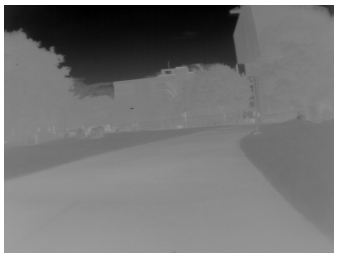
(l) 3.00pm - thermal



(c) 7.00am - visible

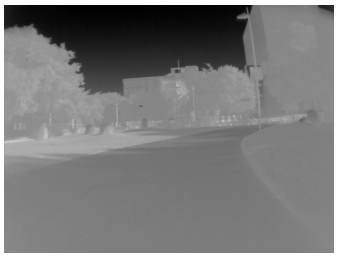
(d) 7.00am - thermal

(m) 5.00pm - visible



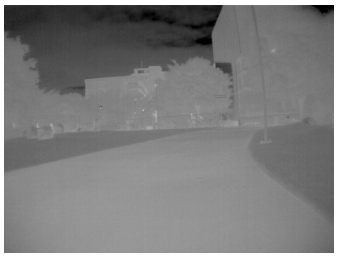
(n) 5.00pm - thermal

(e) 9.00am - visible



(f) 9.00am - thermal

(o) 7.00pm - visible



(p) 7.00pm - thermal

(g) 11.00am - visible

(h) 11.00am - thermal

(q) 9.00pm - visible

(r) 9.00pm - thermal

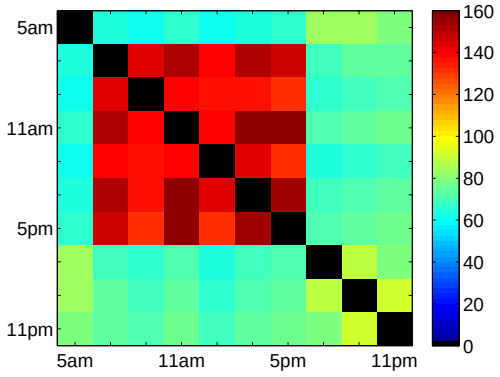
(i) 1.00pm - visible

(j) 1.00pm - thermal

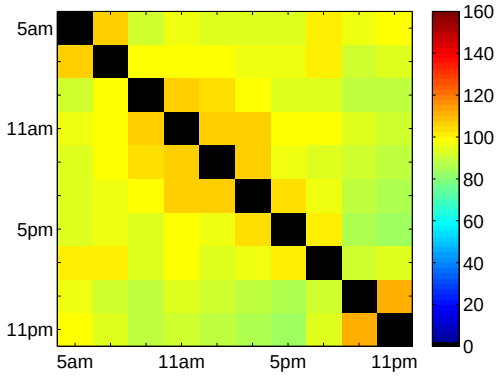
(s) 11.00pm - visible

(t) 11.00pm - thermal

Fig. 3. Dataset snapshots captured from the same location across a 20 hour period. The extreme change in appearance in the visible spectrum is clearly apparent across the whole day. The thermal-infrared modality exhibits lower contrast but remains more consistent over time.



(a) visible



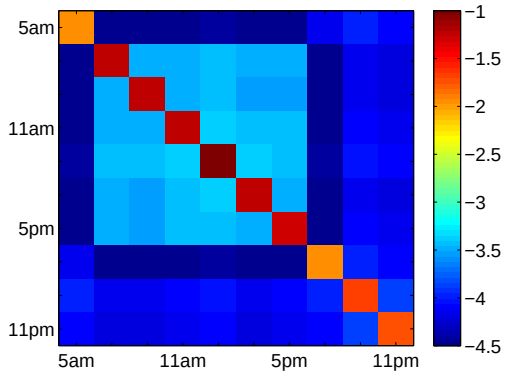
(b) thermal

Fig. 4. Average word overlap between different times of day.

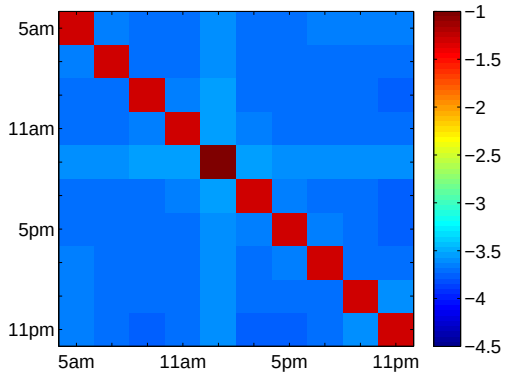
V. Results

Figure 4 shows the average quantity of words per frame which re-occur between matching frames for all possible pairs of times. Poor word overlap occurs between dark times (before dawn or after dusk) in the visible modality, and between dark times and well-lit times. In some cases this is due to a low absolute number of detected features (as low as 20 in some images), but primarily The thermal-infrared modality does not experience the same sharp drop-off outside of daylight hours, but instead exhibits a gradual decrease in word overlap as the time difference between datasets increased. However, the total number of overlapping words for the visual modality exceeds the word overlap for the thermal-infrared modality during daylight hours. Promisingly, superior word overlap is achieved in thermal-infrared for pairs of times in which the visible modality is weak.

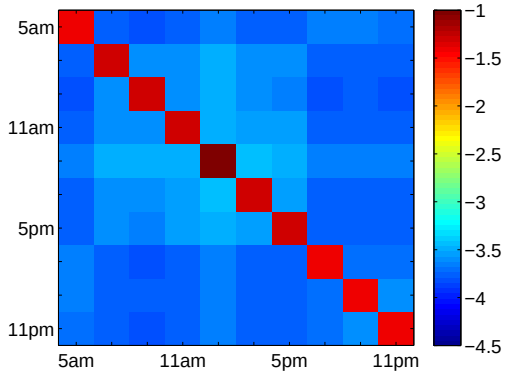
Figure 5 shows the average normalised log observation likelihood across all locations between different times of day for the visible, thermal-infrared and combined FAB-MAP implementations. A similar pattern to the word overlap appears between the visible and thermal-infrared modalities; the visible modality provides good performance during day-light hours but poor performance during night-time, whereas the thermal-infrared modality provides consistent performance



(a) visible



(b) thermal



(c) combined

Fig. 5. Average observation log-likelihood between different times of day.

which gracefully degrades with increasing difference in time between route traversals. The effects of training the Chow-Liu tree on the data collected at 1pm is manifested by the slight increase in observation likelihood matches to the 1pm route traversal (since the Chow-Liu tree provides the most accurate model of the data used to generate it).

The combined representation provides the best overall observation likelihood results; the thermal-infrared modality provides robustness to extreme changes in time between route

traversals, and the visible modality provides superior recall to locations revisited during daylight hours.

The route map shown in Figure 1 has been colored to reflect the observation likelihood of each modality in recognizing each location between the 5am and 7am dataset. Large stretches of the route were poorly lit at 5am, such as much of the parkland, and were poorly recognized by the visual modality. However, the thermal-infrared modality provides consistently high observation likelihood across the full length of the dataset.

VI. Conclusion

The results for word overlap indicate that the thermal-infrared images are more temporally robust under the day-night cycle for the purpose of loop closure. The visible-spectrum images share very few visual words before 7am and after 5pm, corresponding to night-time illumination, whereas the thermal-infrared images demonstrate word overlap from the full period from 5am to 9pm.

The average observation log-likelihoods demonstrate the increased reliability of the thermal-infrared modality for the purpose of loop closure detection across a full day. While the visible modality provides higher match likelihoods during daylight hours, the thermal-infrared modality is capable of consistently matching locations between extreme changes in ambient lighting, such as between the 5am and 7am route traversals.

By combining the visual words from both modalities and building a Chow-Liu tree to caputre inter- and intra- modality dependencies, the observation likelihood performance is further increased, demonstrating how the complementary nature of these modalities can be exploited to boost loop closure performance across a full day.

References

[1] M. Cummins and P. Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *The International Journal of Robotics Research*, 27(6):647–665, June 2008. 1, 2

[2] K Iagnemma, M Buehler, and S Singh. Special issue on the 2007 DARPA urban challenge. *Journal of Field Robotics*, 25:423–860, 2008. 1

[3] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S. Kammel, J. Z. Kolter, D. Langer, O. Pink, V. Pratt, et al. Towards fully autonomous driving: Systems and algorithms. In *Intelligent Vehicles Symposium (IV)*, pages 163–168, 2011. 1

[4] C. McManus, P. T. Furgale, B. Stenning, and T. D. Barfoot. Lighting-Invariant Visual Teach and Repeat Using Appearance-Based Lidar. *Journal of Field Robotics*, 2013. Manuscript ID: ROB-12-0037, submitted on May 17th, 2012. 1

[5] G. Dubbelman, W. van der Mark, J. C. van den Heuvel, and F. C. A. Groen. Obstacle detection during day and night conditions using stereo vision. In *Intelligent Robots and Systems. IEEE/RSJ International Conference on*, pages 109–116, 2007. 1

[6] D. G Lowe. Object recognition from local scale-invariant features. In *Computer Vision. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157, 1999. 1

[7] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008. 1, 3

[8] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth. FAB-MAP+ RatSLAM: Appearance-based slam for multiple times of day. In *Robotics and Automation (ICRA), IEEE International Conference on*, pages 3507–3512, 2010. 1

[9] C. Valgren and A. Lilienthal. SIFT, SURF and seasons: Long-term outdoor localization using local features. In *Proceedings of the European Conference on Mobile Robots (ECMR07)*, 2007. 1

[10] S. El-Tawab, M. Abuelela, and Y. Gongjun. Real-time weather notification system using intelligent vehicles and smart sensors. In *Mobile Adhoc and Sensor Systems. MASS. IEEE 6th International Conference on*, pages 627–632, 2009. 1

[11] M. Felton, K. P. Gurton, J. L. Pezzaniti, D. B. Chenault, and L. E. Roth. Comparison of the Inversion Periods for Mid-wave IR (MidIR) and Long-wave IR (LWIR) Polarimetric and Conventional Thermal Imagery. Technical report, DTIC Document, 2010. 1

[12] S. Vidas, R. Lakemond, S. Denman, C. Fookes, S. Sridharan, and T. Wark. An exploration of feature detector performance in the thermal-infrared modality. In *Digital Image Computing Techniques and Applications (DICTA), International Conference on*, pages 217–224, 2011. 2

[13] Karel Zuiderveld. Graphics gems IV. pages 474–485. Academic Press Professional, Inc., San Diego, CA, USA, 1994. 2, 3

[14] Z. Jia, H. Wang, R. Caballero, Z. Xiong, J. Zhao, and A. Finn. Real-time content adaptive contrast enhancement for see-through fog and rain. In *Acoustics Speech and Signal Processing (ICASSP), IEEE International Conference on*, pages 1378–1381, 2010. 2

[15] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477, 2003. 2

[16] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3):462–467, 1968. 2, 3