# Kinetics and Scene Features for Intent Detection

Raksha Ramesh [1], Vishal Anand [1], Ziyin Wang [1], Tianle Zhu [1], Wenfeng Lyu [1], Serena Yuan [1], and Ching-Yung Lin [1,2]
[1] {rn2486, va2361, zw2605, tz2434, wl2733, sy2657, c.lin}@columbia.edu; [2] {cylin}@graphen.ai
[1] Columbia University, New York, NY, USA
[2] Graphen AI, New York, NY, USA

## ABSTRACT

We create multi-modal fusion models to predict relational classes within entities in free-form inputs such as unseen movies. Our approach identifies information rich features within individual sources – emotion, text-attention, age, gender, and contextual background object tracking. These information are absorbed and contrasted from baseline fusion architectures [1]. These five models then showcase future research areas on this challenging problem of relational knowledge extraction from movies and free-form multimodal input sources. We find that, generally, the Kinetics model added with Attributes and Objects beat the baseline models.

## CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Information extraction**; **Computer vision**; **Scene understanding**; **Activity recognition and understanding**; *Discourse, dialogue and pragmatics*; *Lexical semantics*; *Knowledge representation and reasoning*; *Image representations*; *Object recognition*; Matching.

## KEYWORDS

natural language processing, computer vision, multi-modal fusion, information extraction, neural networks, scene detection, object recognition, video understanding, activity recognition

**ACM Reference Format:**
Raksha Ramesh, Vishal Anand, Ziyin Wang, Tianle Zhu, Wen-feng Lyu, Serena Yuan, and Ching-Yung Lin. 2020. Kinetics and Scene Features for Intent Detection. In *Utrecht '20: ACM International Conference on Multimodel Interaction, October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3395035.3425641

## 1 INTRODUCTION

Recent progress in neural networks and deep learning shows exciting potential for the hard problem on video understanding on free-form inputs. Today's digital contents are inherently multi-modal. Promising results achieved in deep analysis tasks on image, speech, audio, text, and video domains have motivated processes to learn video representations to better exploit abundant multimodal clues for video annotation using natural language and video question answering. However, there is still a knowledge limit on computer vision systems to develop deep analysis in semantic relation understanding among multimodal entities. Most computer vision systems analyzing long duration videos taking one or more modalities fail to effectively condense the extracted sources to answer natural language processing or knowledge graph queries.


*ICMI '20 Companion, October 25–29, 2020, Virtual event, Netherlands*

ACM ISBN 978-1-4503-8002-7/20/10…$15.00
https://doi.org/10.1145/3395035.3425641

## 2 BACKGROUND AND RELATED WORK

A multimodal model can represent the joint representations of multiple input modalities [27]. Different modalities are characterized by different statistical properties. Moreover, different modalities with different statistical properties will influence over the prediction output.

Multimodal model is widely used in many areas such as medical [26] [29] [18], education [8] [5] [20], transportation [16] [25] [6] and media [17] [28] [21]. It has been applied to a broad set of applications such as classification and information retrieval tasks. The multimodal model is useful for classification such as text classification, video classification, sentiment classification. Krishnamurthy et al. [13] combined the audio, video, and text features to detect deception in videos. Huang et al. [9] used fusion-based multimodal attention model to exploit the internal correlation between visual and textual features for joint sentiment classification.

Recognizing fine-grained social relationships & actions from small-scale datasets is a challenging task that requires us to leverage transfer learning to boost the generalisability of the model. In this paper we extend the recent work [1] on semantic relationship understanding by learning holistic scene and text representations. We present a discussion on the most useful feature sets to model the target task of pairwise relation prediction which can be useful for other video understanding tasks such as visual question answering that involve multimodal sources.

## 3 METHODOLOGY

Firstly, we process the video files by capturing changes in the story line through scene boundary detection. This enables us to better localize the entities of interest. We then detect and map each entity with the associated ground-truth through face clustering and SIFT based object mapping. Next, we extract multiple visual and contextual object cues to enrich the scene modality and extract embeddings for speaker-diarized text. We then train a unified model to produce probabilities for a pairwise entity relationship.

In the following section, we present a brief overview of the building blocks that is constructed on top of baseline pipeline [1].

## 4 DATASET

The High-Level Video Understanding (HLVU) dataset [4] includes 10 movies that are suitable for researching the relationship between

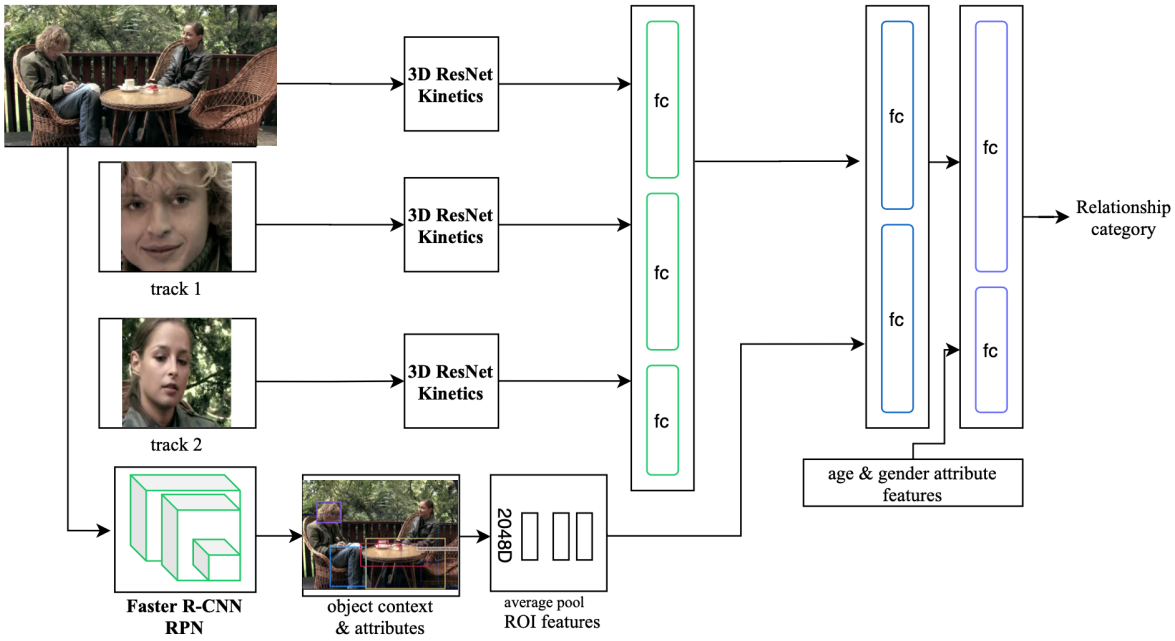



**Figure 1: Architecture of the multi-modal fusion model**

entities. The HLVU dataset meets the important requirements for selecting movies such as the duration of the movies (different lengths of movies: 6 long movies and 4 short movies), the quality of the video, and the clarity of the storyline. The training set is annotated with images of key characters, locations, objects and concepts along with pairwise relations between the entities. There are sixty symmetric relationships to learn. The testing sets only contain the images of key entities.

## 5 BUILDING BLOCKS

- **Scene Detection**:
  We group together multiple adjacent shots that are semantically related into scenes to capture changes in storyline. We adopt the dynamic programming approach in [23] and use HSV color histograms as features to predict scene boundaries.
- **Face and Object Entity mapping** :
  We densely detect faces in each video frame with dlib's face detector and facial landmark predictor. The aligned faces are mapped to 128D embeddings with dlib's ResNet trained with triplet loss. We use Chinese whispers [3] for clustering unidentified faces and use KNN to assign each cluster to the majority label predicted.
  The object and location entities are localized in the scenes through SIFT based feature matching [24]. The best matches between the template and the scene are computed by sorting the feature descriptors distances obtained from keypoints.
- **Speaker Diarization and Face Mapping** :
  We use CMU-Sphinx to generate speaker-separated audio splits and perform speaker identification using scene based face mapping. The speakers are identified by using dlib's [11] 68 point facial landmarks to capture the movement of the speaker's lips and lip motion is estimated using the relative ratio of vertical lip shape to face size.

## 6 APPROACH

### 6.1 Kinetics Features

3D CNNs utilize spatio-temporal convolutional kernels that are best suited for action recognition tasks. The Kinetics dataset contains 400 human action categories spanning singular person actions, and interactions with other persons and objects. To capture scene dynamics, we therefore extract features for video sequences within a scene from 3D ResNet network [7] pre-trained on the Kinetics dataset [10].

### 6.2 Attribute Features for Person Entities

We automatically extract age and gender attributes for every person entity of interest. We follow the stacked CNN implementation as per [14] trained on the Aidence benchmark.

### 6.3 Emotion Features for Person Entities

We explore the association of emotions expressed between person entities in determining their social relationship through two modalities - facial expressions and speech emotions. For facial emotion recognition, we fine-tune the VGGFace network [19] on the FER2013 emotion dataset to detect six emotion categories - happy, sad, anger, disgust, fear and surprise. We extract emotion embeddings from the fine-tuned model for every person entity for a sequence of 16 frames within a scene.

### 6.4 Text Features

First, we use CMU-Sphinx to produce speaker-separated audio splits and then use Google-API with speaker identification to assign speakers with names. After that, we fine-tune the DialogRE [30] model, which is based on bert, to extract relations from our dialogues. With automated-mapping of 36 relation-categories to our 60 relation-categories, we augment our data. Then we take the results from DialogRE model as our text features to train with kinetics features and attributes features.

### 6.5 Contextual Background Objects Features

We extract image embeddings from Faster R-CNN network with a Resnet-101 backbone. [2]. The object proposals are pre-trained on Visual Genome data [12] which contain visual attributes like colours and clothing along with densely annotated objects.

Anchor boxes of different scales and aspect ratios are extracted from the Faster R-CNN's [22] region proposal network through selective search. These proposals are further refined through non-maximum suppression. We threshold the number of candidate object regions to 10 and filter out proposals with weak confidence scores. We then extract an image embedding of dimension 2048 to represent each proposal. The features are average pooled and concatenated with global kinetics features above. These features from the salient image regions embed contextual cues from the surroundings of the co-located entities and aid in better scene representations.

## 7 IMPLEMENTATION

We first localize entities and create individual scene-level tracks for all entities that occur in a scene. The tracks Figure 1 are temporally aligned with the original scene and contain cropped bounding boxes of an entity's body regions obtained through SSD [15]. We further choose contiguous sequences of 16 frames to represent an interaction between the entities and use an appropriate frame margin to limit the search space for sequences. We initialize our network with weights from a 3D ResNet network pre-trained on the Kinetics dataset and further fine-tune the model with additional fully-connected layers to learn the primary sixty relationship categories in the dataset.

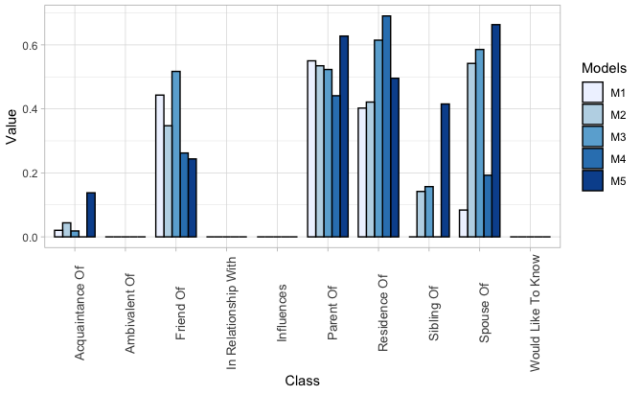
We experiment with fusing embeddings obtained from facial expressions, conversational dialogues and also inject age and gender attribute features into the network and jointly train the network. Further, we explore the role of contextual cues from background objects by fusing the feature maps from Faster-RCNN's object proposals with the Kinetics' features.

For training, we use cross-entropy loss with Adam optimizer. We choose initial learning rate of 0.001 and reduce it by a factor of 10 when validation loss saturates. We also use batch normalization and dropout for regularization.

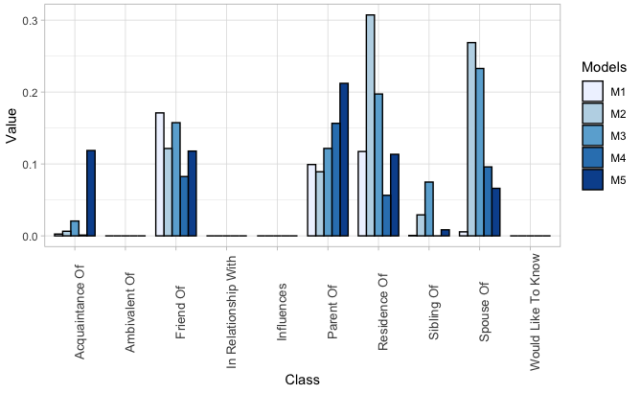
In the next section we analyse the contribution of all these feature sets towards the target task of semantic relation understanding.

## 8 EVALUATION AND RESULTS

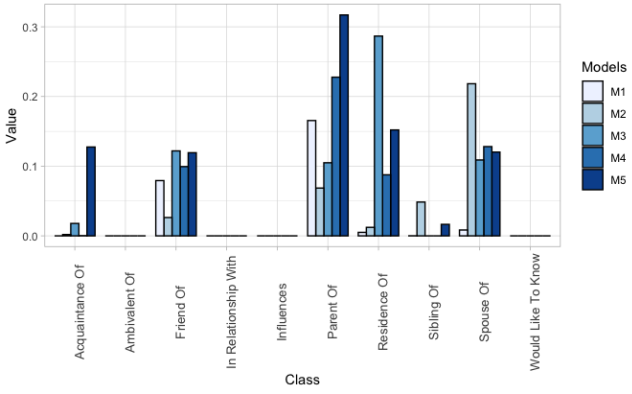
In our study, we compare and contrast the performance of 5 models, namely *Kinetics* Model, *Kinetics with Age and Gender* Model, *Kinetics with Age, Gender and Emotion* Model, *Kinetics with Age, Gender and Text* Model and *Kinetics with Age, Gender and Object context* Model. To evaluate the model we perform cross validation and compare the




**(a) Precision**




**(b) Recall**




**(c) F-1 Scores**

**Figure 2: Multi-Modal Kinetics Fusion Scores for Top-k classes, k=10**
**M1 = Kinetics features**
**M2 = Kinetics + Attributes (Age + Gender)**
**M3 = Kinetics + Attributes (Age + Gender) + Emotion**
**M4 = Kinetics + Attributes(Age + Gender) + Text**
**M5 = Kinetics + Attributes(Age + Gender) + Object Context**

| Classes | M1 | | | M2 | | | M3 | | | M4 | | | M5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Parent Of | 0.51 | 0.08 | 0.14 | 0.47 | 0.05 | 0.09 | 0.46 | 0.07 | 0.12 | 0.62 | 0.22 | 0.32 | 0.67 | 0.23 | 0.34 |
| Sibling Of | 0 | 0 | 0 | 0.19 | 0.04 | 0.06 | 0.21 | 0.1 | 0.13 | 0 | 0 | 0 | 0.55 | 0.01 | 0.02 |
| Spouse Of | 0 | 0 | 0 | 0.39 | 0.25 | 0.30 | 0.46 | 0.23 | 0.30 | 0 | 0 | 0 | 0.53 | 0.05 | 0.1 |
| Friend Of | 0.01 | 0.07 | 0.02 | 0.01 | 0.09 | 0.01 | 0 | 0.05 | 0.01 | 0.04 | 0.62 | 0.08 | 0.01 | 0.1 | 0.02 |

**Table 1: Comparison of precision, recall and F1 scores for four prominent classes for the movie *Spiritual Contact***

precision, recall and F1 scores across the five models. We analyze the contribution of embedding the age gender attributes (M2), emotion (M3), text (M4) and object cues (M5) to the baseline model (Kinetics features only) (M1) independently.

Table 1 summarizes the scores for the top four classes present in the evaluation movie *Spiritual Contact* across all the models.

We further extend this analysis by averaging the scores across all training movies. The plots in Figure 2 visualize the weighted average precision, recall and F1 scores across the six training and validation movies. Some key observations from the plots in Figure 2 and the Table 1 are summarized below:

- The precision, recall and F1 scores increase significantly for the "Sibling Of" and "Spouse Of" classes when the age and gender attribute features are embedded into the baseline model. We further see an increase in precision for these classes when emotion is embedded with the age & gender attributes. For example, in *Spiritual Contact* (Table 1), we observe a percentage increase of 10.5% and 18% in precision for these classes.
- From Table 1, we can observe, age & gender attributes is not a discriminatory feature for "Friend Of" class. But it is also interesting to note that while these attributes alone do not contribute, embedding emotion increases the scores of this class on average as seen in Figure 2.
- In Table 1, for "Parent Of" class, precision of M5 increases by 31% from the baseline. This trend is true across all movies. Therefore, introducing background object features from the scene increases the scores on the "Parent Of" classes which suggests the model recognizes certain visual contextual cues to be indicative of parent-child relationships. Similarly, we also observe higher recall and F1 scores for M4 in Table 1 and higher average recall and F1 scores in Figure 2 in comparison to M1, M2 and M3. This indicates the conversational cues are more effective than emotion to identify parent-child relationships.
- The low precision and recall scores on some classes like "Influences", "In Relationship With" and "Would Like to Know" are due to a lack of training samples. Although these classes are among the top ten classes present in the training set i.e., contain most number of samples, the video sequences we extract for these classes either belong only to one movie within the training set, or to particular pairs of entities within that movie. This can be attributed to the longer duration of these movies (leading to the sampling of more sequences) relative to other movies, resulting in poorer generalisability across all the other movies. Further, among the 60 relationship categories mentioned in the dataset, we are able to extract video sequences for 30 relationship categories. This suggests the need to augment the existing dataset to increase performance on other classes.

In summary, while adding just age and gender attributes can result in some irregularities in the scores, further embedding emotion, text or object context in the model significantly improves the scores across most classes.

## 9 CONCLUSION AND FUTURE WORK

Human computer interaction focus looks very promising as the path-forward in deep-video-understanding and information extraction from free-form multi-modal inputs. When attribute features of age and gender, along with emotion, text or object context are added to the baseline model, our metrics noticeably improve. This suggests the ability of feature sets effectively encapsulating semantic relations between entities in multi-modal sources. The features that perform better have a larger sample size, and we find majority of data sources are skewed due to imbalanced training classes, that can be augmented for better training and human evaluation.

A future direction of research could be to incorporate head tracking to improve face-entity mapping and incorporate crisper speaker diarization to improve the results on the text fusion model. A higher focus on human centered interaction metrics of humor, anger or passion in text and visual cues may lead to higher scores and deeper understanding of videos.

## REFERENCES


[1] Vishal Anand, Raksha Ramesh, Ziyin Wang, Yijing Feng, Jiana Feng, Wenfeng Lyu, Tianle Zhu, Serena Yuan, and Ching-Yung Lin. 2020. Story Semantic Relationships from Multimodal Cognitions. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. https://doi.org/10.1145/3394171.3416305

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[3] Chris Biemann. 2006. Chinese whispers-an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*. 73–80.

[4] Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. 2020. HLVU: A New Challenge to Test Deep Understanding of Movies the Way Humans do. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 355–361.

[5] Kristina Danielsson and Staffan Selander. 2016. Reading Multimodal Texts for Learning–A Model for Cultivating Multimodal Literacy. *Designs for learning* 8, 1 (2016), 25–36.

[6] Sune Djurhuus, Henning Sten Hansen, Mette Aadahl, and Charlotte Glümer. 2016. Building a multimodal network and determining individual accessibility by public transportation. *Environment and Planning B: Planning and Design* 43, 1 (2016), 210–227.

[7] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6546–6555.

[8] Dawnene D Hassett and Jen Scott Curwood. 2009. Theories and practices of multimodal education: The instructional dynamics of picture books and primary classrooms. *The Reading Teacher* 63, 4 (2009), 270–282.

[9] Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, Jie Xu, and Zhoujun Li. 2019. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems* 167 (2019), 26–37.

[10] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[11] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research* 10 (2009), 1755–1758.

[12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123, 1 (2017), 32–73.

[13] Gangeshwar Krishnamurthy, Navonil Majumder, Soujanya Poria, and Erik Cambria. 2018. A deep learning approach for multimodal deception detection. *arXiv preprint arXiv:1803.00344* (2018).

[14] Gil Levi and Tal Hassner. 2015. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 34–42.

[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[16] Meifeng Luo and Thomas A Grigalunas. 2003. A spatial-economic multimodal transportation simulation model for US coastal container ports. *Maritime Economics & Logistics* 5, 2 (2003), 158–178.

[17] Pradeep Natarajan, Shuang Wu, Shiv Vitaladevuni, Xiaodan Zhuang, Stavros Tsakalidis, Unsang Park, Rohit Prasad, and Premkumar Natarajan. 2012. Multimodal feature fusion for robust event detection in web videos. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1298–1305.

[18] Lei Pang, Shiai Zhu, and Chong-Wah Ngo. 2015. Deep multimodal learning for affective analysis and retrieval. *IEEE Transactions on Multimedia* 17, 11 (2015), 2008–2020.

[19] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. (2015).

[20] Anthony G Picciano. 2017. Theories and frameworks for online education: Seeking an integrated model. *Online Learning* 21, 3 (2017), 166–190.

[21] Dimitrios Rafailidis, Pavlos Kefalas, and Yannis Manolopoulos. 2017. Preference dynamics with multimodal user-item interactions in social media recommendation. *Expert Systems with Applications* 74 (2017), 11–18.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[23] Daniel Rotman, Dror Porat, and Gal Ashour. 2016. Robust and efficient video scene detection using optimal sequential grouping. In *2016 IEEE International Symposium on Multimedia (ISM)*. IEEE, 275–280.

[24] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision*. Ieee, 2564–2571.

[25] Young Joon Seo, Feilong Chen, and Sae Yeon Roh. 2017. Multimodal transportation: The case of laptop from Chongqing in China to Rotterdam in Europe. *The Asian Journal of Shipping and Logistics* 33, 3 (2017), 155–165.

[26] Rajiv Singh and Ashish Khare. 2014. Fusion of multimodal medical images using Daubechies complex wavelet transform–A multiresolution approach. *Information fusion* 19 (2014), 49–60.

[27] Nitish Srivastava and Russ R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*. 2222–2230.

[28] Di Wang, Xinbo Gao, Xiumei Wang, and Lihuo He. 2015. Semantic topic multimodal hashing for cross-media retrieval. In *Twenty-fourth international joint conference on artificial intelligence*.

[29] Xinzheng Xu, Dong Shan, Guanying Wang, and Xiangying Jiang. 2016. Multimodal medical image fusion using PCNN optimized by the QPSO algorithm. *Applied Soft Computing* 46 (2016), 588–595.

[30] Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-Based Relation Extraction. *arXiv preprint arXiv:2004.08056* (2020).