VISION-BASED ACTION RECOGNITION IN THE INTERNAL CONSTRUCTION SITE USING INTERACTIONS BETWEEN WORKER ACTIONS AND CONSTRUCTION OBJECTS

J. Y. Kim and *C. H. Caldas
University of Texas at Austin
301 East Dean Keeton
Austin, TX, 78712, USA
*(*Corresponding author: caldas@mail.utexas.edu)*

VISION-BASED ACTION RECOGNITION IN THE INTERNAL CONSTRUCTION SITE USING INTERACTIONS BETWEEN WORKER ACTIONS AND CONSTRUCTION OBJECTS

ABSTRACT

This paper presents a novel action recognition method for observing human workers using interactions between actions and related objects on an internal construction site. This method can be used to measure work rates for labour productivity monitoring. This monitoring is critical because the performance of a construction project is significantly impacted by labour productivity. However, construction sites are generally crowded with a large number of workers and objects. Such congestion disrupts the accurate, automatic recognition of construction workers' actions. This congestion is one reason that existing automatic action recognition studies of construction areas mainly focus on workers' actions themselves. However, the crowded conditions mean that sites could offer a great deal of clues that could be used for automatic action recognition. According to psychological studies, interactions clearly take place between human actions and related objects, such as between hammering and a hammer. Humans use these interactions to recognize actions or objects more accurately. On the construction site, workers, materials, tools, and equipment are carefully planned out ahead of actual construction. The categories of workers and objects are pre-defined and, as noted, specific interactions define relations between worker actions and objects. In this paper, the interactions are limited to human workers and their hand-held objects. Action recognition results can be combined with hand-held object information to improve recognition accuracy. With the limited interactions, experiments in this paper show a significant improvement in action recognition. This paper describes the utilization of these interactions to improve construction action recognition accuracy based on human skeleton data and 2D color video from Microsoft KINECT sensor.

KEYWORDS

Action Recognition, Work Sampling, Activity Analysis, KINECT, Skeleton, Human-Object Interaction

INTRODUCTION

The overall productivity of construction projects is highly dependent on the productivity of the labour force. Any construction project can be broken into a series of single laborer tasks. Such dependency on labour originates from the uniqueness of each construction project. The unique projects call for a number of modifications to their construction process. This modification requires a wide use of human labour (Allmon, 2000). Therefore, for construction projects to succeed, it is critical that construction labour achieve efficient productivity.

Improving labor productivity requires measurement. Measuring productivity is necessarily preceded by gathering on-site productivity data (Gong & Caldas, 2010). In the United States; however, 31% of contractors surveyed have no formal process for measuring, monitoring, or documenting construction productivity (Motwani et al., 1995). This lack of productivity measurement can be explained by the difficulties and time lags involved in measuring at the project level (Oglesby, 1989). Measuring methods include an array of project-level information systems, direct observation methods, and survey/interview based methods (Gong & Caldas, 2010).

Some researchers have recommended work sampling as a tool for measuring sources of inefficiency (Oglesby, 1989; Hanna, 2010) causing low labour productivity. However, the benefits of work sampling are offset by its tedious and time-consuming (Liju & Koshy, 2011) manual observation. Moreover, to attain statistical significance, it requires a large amount of observations. For instance, for a site with 500 craft workers to have 95% statistical significance, 253 observations (51%) should be made per hour (CII, 2010). Furthermore, if they notice that they are under observation, construction workers may alter their behavior. Therefore, an automatic and objective observation method is required to accurately monitor labour productivity on a construction site.

BACKGROUND

Automatic Human Action Recognition in Construction

One of the most direct methods of recognizing action is to attach sensors to human workers. To detect masonry actions, Joshua and Varghese (2011) used accelerometers which they attached to a masonry worker's body. They tested recognition accuracy by varying the positions (left, right waist) of the sensor, signal features, and classifiers. They focused on distinct actions such as fetch and spread mortar, fetch and lay brick, and fill joints. Alwasel and Haas (2011) developed a joint angle sensor system to detect work-related musculoskeletal disorders (WMSDs). The main components of the system were a permanent magnet as a magnetic field source and a sensing element. The magnet is attached to a moving part, such as an upper arm, and a sensing part is attached to the torso. The sensor system detects the angle of the arm and monitors the worker's musculoskeletal disorders (Alwasel and Haas, 2011).

2D video is a useful tool for human action recognition because it is an inexpensive and easy means of gathering data at the job-site. Furthermore, 2D video-based methods, such as videotaping and time-lapse video, have already been used for productivity analysis on construction jobsites (Gong and Caldas 2010). Peddi et al. (2009) used 2D video to measure productivity in real-time; with a wireless video camera, using the silhouette approach, they were able to extract human skeletons from video and recognize workers' actions. They applied this method to productivity measurement and compared its accuracy with manual observation. 2D video was also used to recognize such actions as transporting, traveling, bending, nailing, and aligning (Gong and Caldas 2011). This approach used scale invariant feature transform (SIFT) descriptor and bag of video feature word approach to recognize those actions. This study was the first method to use scale and rotation invariant image features to recognize worker action on a construction site.

Researchers have also monitored worker health and safety using 3D imaging systems such as a time of flight camera. Gonsalves and Teizer (2009) proposed a tracking algorithm that focused on range or distance information to provide a 3D perspective of the human target (i.e. 3D point clouds). The authors adopted the star skeleton structure to generate a segmented human target (Fujiyoshi, H., & Lipton, 1998). The proposed system classifies different activities by incorporating the use of a particle filter. This system is also able to detect multiple people and track their paths.

Microsoft KINECT provides human skeleton information, video stream (RGB), and point clouds (Depth). The human skeleton is composed of 20 set of human joint coordinates. Escorcia et al. (2012) used skeleton information and bag of pose approach to recognize five distinct drywall installation actions such as caulking, hammering, idling, painting, and walking. Han et al. (2012) developed a KINECT-based unsafe action detection system. They performed a case study with motion data for unsafe action in ladder climbing.

Recently, Cheng et al. (2013) developed a task-level activity analysis system using data fusion of spatio-temporal and workers' thoracic posture data to automatically measure work rates for activity analysis.

Applications of Object-Human Interactions in Computer Vision

Gupta et al. (2009) applied human-object interactions for both object and action recognition. Their model interprets human-object interaction with an object, reach motion, object manipulation motion, and object reaction information. The Bayesian approach was used to integrate this information and to understand the human and object interaction. Another method of action recognition considering human-object interaction was presented by Yao and Fei-Fei (2010a). The authors used action and object information from 2D images. The basic idea was that "recognizing one facilitates the recognition of the other." They infer actions by finding the maximum function value of activity class, object, and human action information. Desai's (2010) the context-based action recognition method (discriminative model) uses action and nearby object information from 2D images to detect action. Yao and Fei-Fei (2010b)

introduced a random field model to encode the mutual context of objects when object and action are considered simultaneously. Objects and human body parts are related; human action and body parts also are related.

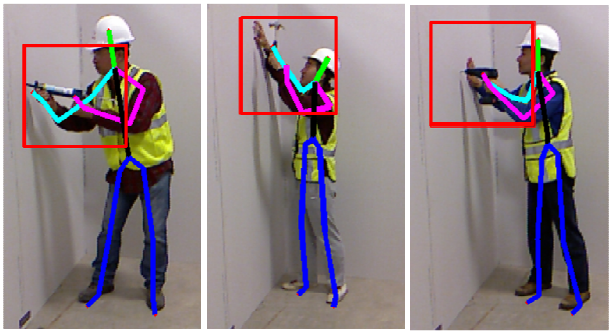
The studies in construction area mainly focused on visually distinct actions such as caulking and walking. Focusing on those distinct actions has the limitation of using such methods because there are actions that are similar but different such as caulking and drilling. However, a construction site includes a large amount of other information to help recognize similar actions. According to psychological studies, humans recognize actions with object information. Gallese et al. (1996) explained that humans can more easily recognize a task if they can see both the action and the object used to perform that action. For example, humans might observe a skeletal representation of a worker's arm moving up and down, and not recognize the specific task without information that the worker is using a hammer. These interactions clearly take place between the construction worker's actions and related objects such as tools and materials. In computer vision area, there are several studies using these human-object interactions. However, their studies were limited to the general actions of daily life, and their data sources were 2d image or video. Therefore, this paper's novel approach is focusing on easily confused construction actions and imitating the human action-recognition process by combining action information with object information. The data sources of this approach are human skeleton information and 2d video from Microsoft KINECT sensor.

OBJECTIVES AND SCOPE

This chapter presents a novel, action-recognition method of observing human workers' interactions between objects and actions related to those objects on an internal construction site. The main objective is to show how action-recognition results can be improved with object information. The basic steps of this method are (1) action recognition, (2) object recognition (i.e. tools or materials), and (3) combination of the two results to improve action recognition. The job selected for this research was drywall installation because it requires 3 easy-to-confused tasks: caulking, hammering, and screwing.

INPUT DATA

The worker's motion-capture device, used in this study is a Microsoft KINECT sensor. This device provided human skeleton information, 2D color video, 3D depth video, and sound. Figure 1 (a) – (c) shows example data of the three actions (i.e. caulking, hammering, and screwing) captured by KINECT. Human skeleton information of the actions is composed of 20 coordinates of human joints such as head, left hand, and right hand. The resolution of the 2D video was 640 x 480 pixels. The human skeleton was used for action recognition and 2D color video was used for object recognition. Depth and sound were not used. The KINECT was implemented by an OpenNI framework and NITE libraries (OpenNI Organization, 2012). The red boxes in each image of Figure 1 are primed areas for detecting hand-held objects. The author gathered training and testing dataset containing the skeleton of workers, and the 2d video of the workers performing the three tasks.



(a) Caulking (b) Hammering (c) Screwing

Figure 1 –Human Skeleton Data for Action Recognition

ACTION RECOGNITION

The four steps in action recognition are shown in Figure 2. The first step is gathering training and testing skeleton data from KINECT. The second step is normalizing the data which is critical because the captured skeleton data came from various individuals of various heights and widths. This normalization guarantees the robustness of action recognition. The third step is reducing the dimensions of skeleton data because the data includes a large number of dimensions (i.e., 20 joint coordinates) and because the Gaussian mixture model (GMM), the classifier used for action recognition, requires fewer dimensions for training and testing. The study employed principal component analysis (PCA) to generate various dimensions of the skeleton. The last step is training and testing the classifier and recognizing actions using the classifier. These steps are simple and straightforward for initial action recognition; however, for this action recognition, better recognition results may be produced by other methods such as bag-of-pose using a spatio-temporal approach (Escorcia et al. 2012). The simpler method is used here to show the effect of a combination of object information for the recognition of easily confused operation actions.

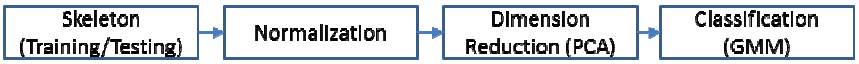


Figure 2 –Action Recognition Steps

Figure 3 shows the results of the simple action-recognition method. The horizontal axis represents the dimensions of the skeleton data, and the vertical axis shows the recognition accuracies for these different dimensions. The results for one dimension of skeleton data were quite low (about 39%); too little information is provided to distinguish actions. With three dimensions, the average accuracy increases to 71%, which is relatively good considering the similarity of the actions and the simplicity of this method. As the number of dimensions continues to increase, however, the average accuracy falls to 25%. With three dimensions, hammering shows the highest average accuracy overall (66%) while caulking shows the best recognition accuracy (71%).

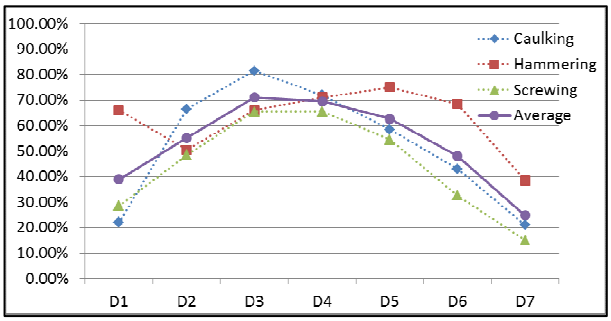


Figure 3 – Initial Action-Recognition Results by Different Numbers of Dimensions for GMM

OBJECT RECOGNITION

Figure 4 shows the five object recognition steps. The first step is obtaining 2D training and testing videos. The second step is extracting hand-held object images from the video by priming original scene using the hand coordinates of skeleton data. Hand-held objects used in this study were a caulking gun, a hammer, and an electronic screw driver. These objects possess direct contextual interactions with the people who handle them. This priming enables us to focus on a specific area and to reduce computation cost. The third step is extracting local features from the primed images. For the feature extraction, SIFT is the one of the most common local feature descriptor. In this study, dense scale-invariant feature transform (D-SIFT) was used because objects in the images were small and vague and because the D-SIFT generates more features than the regular SIFT using those images. The fourth step is forming bags-of-visual-words of individual training and testing images. This approach requires a visual vocabulary which is a set of common features from the feature pool of all the training images. The common features mean the visual words and constitute the visual vocabulary. The bag-of-visual-words is a distribution (or a histogram) of visual word occurrences in each image. Various size of visual vocabulary was applied to create bags-of-visual-words, and their effect on the recognition accuracies was tested. The last step is training and testing

the object classifier. Because it is simple and relatively accurate, the Multiple Naïve-Bayes classifier was applied here.

Figure 4 –Object Recognition Steps

Figure 5 illustrates the object recognition results using the five steps. The horizontal axis is the number of types of vocabulary used for bag-of-visual-words formation, and the vertical axis represents the recognition accuracies by the different types of vocabulary. Overall, a higher vocabulary shows slightly better results, but the accuracy drops off after the number reaches 70. This object-recognition approach has the highest accuracy when the bag-of-visual-words formation uses 60 types of vocabulary. Its average accuracy was 85% (caulking gun 91%, hammer 72 %, and electronic screw driver 91%).

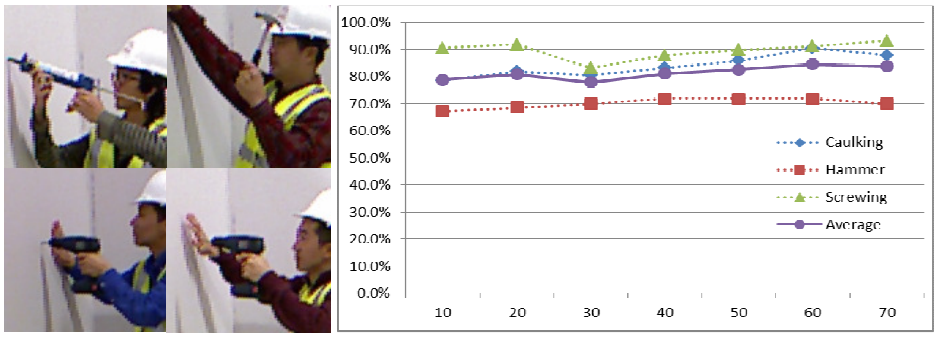


Figure 5 – Objects and Their Recognition Results by Different Numbers of Vocabulary

COMBINATION OF ACTION AND OBJECT RECOGNITION

To combine action with object recognition, actions were recognized first and then objects in the corresponding images of the actions were detected. The weight-based combination method used is shown in Equation 1 (Marszalek et al., 2009), and was originally designed for action recognition of given scenes. The author modified the method to combine action and object probabilities, as follows:

$$g'_a(x) = g_a(x) + \tau \cdot W \cdot g_o(x) \quad (1)$$

The combined probability of action and object is denoted as $g'_a(x)$. The initial action probability is $g'_a(x)$, and the object probability is $g'_o(x)$. The parameter τ is decided by experiments, and weight W is calculated from the object probabilities of training data ($P(Object \mid Image)$).

Figure 6 shows the results of the combination. Significant improvements in action recognition were observed. The horizontal axis is the number of skeleton dimensions used for action recognition, and the vertical axis is the action-recognition accuracy combined with object recognition results. The first to the fourth data series are given as average recognition accuracies of caulking, hammering, and screwing. The accuracies vary with the values of parameter τ. The bottom data series is the initial average action-recognition results. The improvements range from 16 to 66% depending on parameter τ, and the number of skeleton dimensions. Although Parameter τ is set experimentally, the results are not very sensitive when its values range from 0.1 to 2.0. In nearly all cases, the combination results were higher than the individual action recognition and object recognition. This can be explained by the fact that combining object recognition results reinforces the object's related action recognition result and thereby helps distinguish it from other similar, easy-to-be-confused-with actions.

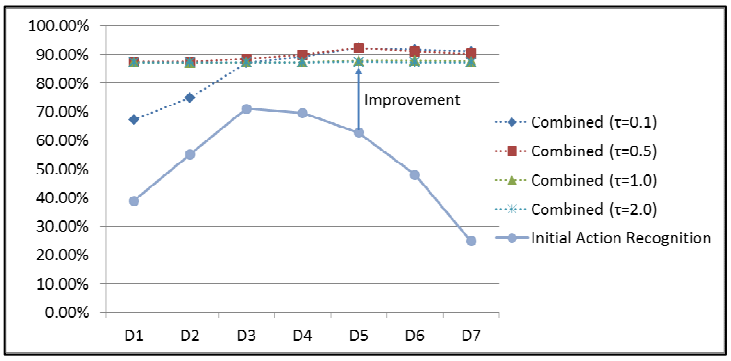


Figure 6 – Average Accuracy Comparisons and Improvements

The combined action recognition accuracy is significantly higher than initial action recognition. The maximum average accuracy improved is 92.2% (when parameter τ is 0.5, skeleton dimensions are 5, and number of visual vocabulary are 60). Table 1 shows numbers for the case of maximum average accuracy. The improved action recognition results were 92.5% for caulking, 88.5 % for hammering, and 95.5% for screwing. Figure 7 is a confusion matrix of this case. The matrix shows the actual performance of the combination. In the matrix, hammering has a relatively lower accuracy than the other two actions.

Table 1 – The Best Combination Results (τ = 0.5, Dimensions = 5, Vocabulary = 60)

Actions	Initial Action Results	Object Results	Combination	Improvement
Caulking	58.5%	87.5%	**92.5%**	**34.0%**
Hammering	75.0%	78.5%	**88.5%**	**13.5%**
Screwing	54.5%	96.0%	**95.5%**	**41.0%**
Average	62.7%	87.3%	**92.2%**	**29.5%**

Caulking	**0.925**	**0.040**	**0.030**
Hammering	**0.025**	**0.885**	**0.015**
Screwing	**0.050**	**0.075**	**0.955**
	Caulking	Hammering	Screwing

Figure 7 – Confusion Matrix of the Best Combination (τ = 0.5, Dimensions = 5, Vocabulary = 60)

CONCLUSIONS

This paper presents a method that combines action and object recognition results, confirming that object recognition improves action recognition (i.e. maximum 92.2 % average accuracy with 29.5% improvement). The authors collect training and test dataset of similar construction actions using a KINECT sensor. With the data, initial actions were recognized by a simple method, then, objects in the corresponding actions were detected. The results were combined together. Results of experiments with the variations of parameters and conditions show that object information significantly increases action recognition. Our findings suggest that object recognition reinforces related action recognition results and thereby is to distinguish one action from similar, easy-to-be-confused-with actions. This logic can be applied to object recognition with action information and can also be generalized for the other combinations with other information. However, this study was performed in a controlled and limited environment. Therefore, verification on a real construction site is required and will be performed.

REFERENCES

Alwasel, A., Elrayes, K., Abdel-Rahman, E. M., & Haas, C. (2011). Sensing Construction Work-Related Musculoskeletal Disorders (WMSDs). *2011 ISARC Proceedings* (pp. 164-169).

CII (Construction Industry Institute). (2010). *Activity analysis guide. Implementation Resource 252–3*. CII.

Desai, C., Ramanan, D., & Fowlkes, C. (2010, June). Discriminative models for static human-object interactions. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 9 –16). 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). doi:10.1109/CVPRW.2010.5543176

Escorcia, V., Dávila, M. A., Golparvar-Fard, M., & Niebles, J. C. (2012, May). Automated Vision-Based Recognition of Construction Worker Actions for Building Interior Construction Operations Using RGBD Cameras. In *Construction Research Congress 2012* (pp. 879–888). American Society of Civil Engineers. Retrieved from http://ascelibrary.org/doi/abs/10.1061/9780784412329.089

Fujiyoshi, H., & Lipton, A. J. (1998, October). Real-time human motion analysis by image skeletonization. In *, Fourth IEEE Workshop on Applications of Computer Vision, 1998. WACV '98. Proceedings* (pp. 15 – 21). Presented at the , Fourth IEEE Workshop on Applications of Computer Vision, 1998. WACV '98. Proceedings. doi:10.1109/ACV.1998.732852

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*(2), 593–609. doi:10.1093/brain/119.2.593

Gong, J., & Caldas, C. (2010). Computer Vision-Based Video Interpretation Model for Automated Productivity Analysis of Construction Operations. *Journal of Computing in Civil Engineering*, *24*(3), 252–263. doi:10.1061/(ASCE)CP.1943-5487.0000027

Gonsalves, R., & Teizer, J. (2009, June). Human Motion Analysis Using 3D Range Imaging Technology. 2009 *ISARC Proceedings*.

Gupta, A., Kembhavi, A., & Davis, L. S. (2009). Observing Human-Object Interactions: Using Spatial and Functional Compatibility for Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(10), 1775 –1789. doi:10.1109/TPAMI.2009.83

Hanna, A. S. (2010). *Construction labor productivity management and methods improvement*. Madison, Wis.: Dr. Awad S. Hanna.

Han, S., Lee, S., & Peña-Mora, F. (2012). Vision-Based Detection of Unsafe Actions of a Construction Worker: A Case Study of Ladder Climbing. *Journal of Computing in Civil Engineering*, *0*(ja), null. doi:10.1061/(ASCE)CP.1943-5487.0000279

Joshua, L., & Varghese, K. (2011). Accelerometer-Based Activity Recognition in Construction. *Journal of Computing in Civil Engineering*, *25*(5), 370–379. doi:10.1061/(ASCE)CP.1943-5487.0000097

Liou, F., & Borcherding, J. D. (1986). Work Sampling Can Predict Unit Rate Productivity. *Journal of Construction Engineering and Management*, *112*(1), 90–103. doi:10.1061/(ASCE)0733-9364(1986)112:1(90)

Marszalek, M., Laptev, I., & Schmid, C. (2009, June). Actions in context. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009* (pp. 2929 –2936). IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009. doi:10.1109/CVPR.2009.5206557

Motwani, J., Kumar, A., & Novakoski, M. (1995). Measuring construction productivity: a practical approach. *Work Study*, *44*(8), 18–20. doi:10.1108/00438029510103310

Oglesby, C. H. (1989). *Productivity improvement in construction*. New York: McGraw-Hill.

OpenNI Organization. (2012). "OpenNI Documentation." OpenNI. Retrieved from http://www.openni.org/ 2012.

Peddi, A., Huan, L., Bai, Y., & Kim, S. (2009). Development of Human Pose Analyzing Algorithms for the Determination of Construction Productivity in Real-Time (pp. 11–20). American Society of Civil Engineers. doi:10.1061/41020(339)2

Cheng, T., Teizer, J., Migliaccio, G. C., & Gatti, U. C. (2013). Automated task-level activity analysis through fusion of real time location sensors and worker's thoracic posture data. *Automation in Construction*, *29*(0), 24–39. doi:10.1016/j.autcon.2012.08.003

Thomas B. Moeslund, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding 104, pages 90–126, 2006

Yao, B., & Fei-Fei, L. (2010a, June). Grouplet: A structured image representation for recognizing human and object interactions. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9 –16). Presented at the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/CVPR.2010.5540234

Yao, B., & Fei-Fei, L. (2010b, June). Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 17 –24). Presented at the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/CVPR.2010.5540235