Multispectral analysis for the determination of lycopene concentration in tomatoes

Marcel Mlynarik[1], Gary A. Atkinson[1], Melvyn L. Smith[1], and Khemraj Emrith[2]

[1] Centre for Machine Vision, Bristol Robotics Laboratory, University of the West of England Bristol, BS16 1QY UK
[2] Mohamed bin Zayed University of Artificial Intelligence, Masdar City, Abu Dhabi UAE

Abstract This paper describes a novel computer vision method for the estimation of lycopene concentration in tomatoes using a multispectral imaging approach with up to 15 bands. It is shown that combining intensity measurements at wavelengths from near-infrared to ultraviolet using a neural network model achieved correlation of R^2=0.977 and RMS error=4.63 mg/kg against ground truth lycopene concentration. Our results are comparable or superior to other methods from the literature, which are analysed in detail in the paper. The method can be reproduced with minimal cost and demonstrates the feasibility of the method for industrial application. The main contribution is that a broader range of wavelengths are considered compared to most previous work, with rigorous analysis using a combination of simple regression and artificial neural networks.

Keywords Machine vision, multispectral, lycopene, tomato

1 Introduction

Tomatoes have a vital role in food supply, accounting for 16% of global vegetable[3] production during the last decade [1]. Tomatoes are a rich source of nutrients, including vitamins A and C, lycopene, and potassium. Lycopene is one of the most valuable bio-active compounds in

[3] Tomatoes are technically fruits but often classified as vegetables in a culinary sense.

tomatoes due to a health stimulating carotenoid with antioxidant properties and helps to prevent cardiovascular diseases, cancers, neurodegenerative maladies, and other conditions [2, 3]. With an estimated global annual production of 180 million tonnes [4] tomatoes are the primary natural source of lycopene in our diets. Lycopene content correlates with the maturity of a tomato [5] and is therefore a critical factor in supply chain logistics for optimising harvesting, transportation and storage.

Humans have a natural ability to assess food quality and safety via a simple analysis of the appearance of the tomato in the visible spectrum. The availability of sensors beyond the visible spectrum and progress in computer vision are extending this basic subjective capability, with 1000s of peer reviewed papers featuring keywords "hyperspectral imaging" and "fruit/vegetable/etc" during the last decade. The latest research is aimed at estimation of properties including ripeness, disease and nutritional value [6].

This paper describes a novel non-destructive method for the estimation of lycopene concentration in tomatoes using multispectral data analysis. The main contribution is that a broad range of wavelengths is considered (15 bands between 365nm and 940nm) and rigorously analysed using a combination of simple regression and artificial neural networks. The outputs offer invaluable information for researchers of automated tomato lycopene estimation (or general ripeness/quality estimation using lycopene as a proxy).

2 Related Work

Traditional methods for the precise measurement of lycopene content are high performance liquid chromatography (HPLC), thin layer chromatography (TLC) [7], and spectrophotometric absorbance (SPM) [8]. These chemometric methods have been available for several decades but are time consuming, require hazardous chemicals and destroy the samples.

Non-invasive spectroscopic techniques such as near infrared spectroscopy (NIRS), nuclear magnetic resonance spectroscopy, Raman spectroscopy (RS) and fluorescence spectroscopy are powerful spectroscopic techniques and have been investigated for applications in the

food industry. However, these methods are mostly expensive, are limited to a small number of sample measurement points, and are dedicated for laboratory use only [9,10].

Consequently, computer vision techniques have been explored that deploy reflected or transmitted light to measure lycopene concentration. Some of these methods use the visual spectrum (VIS) in the form of the CIE L*a*b* colour representation. Other methods use multispectral or hyperspectral techniques, often extended to near-infrared (NIR) and/or ultraviolet (UV) wavelengths.

Methods based on the L*a*b* representation of the visual spectrum. Aries et al. [5] achieved a promising logarithmic regression correlation of R^2=0.96 between lycopene and the a* value from a chroma meter, when averaging 14 spots on the equatorial region of tomatoes. Vazques-Cruz et al. [11], used a similar approach with a point spectrophotometer, to obtain linear regression R^2=0.985 using neural networks (NN) with two hidden layers to map intensities of L*, a*, b*, a*/b* and area of vine leaf to lycopene concentration. Ye et al. [12], claim a lower correlation of R^2=0.81, but using a handheld camera and ambient lighting, thus showing promise for realistic low-cost applications. The highest result found in the literature was a correlation between a* and lycopene of R^2=0.985, from Barrios et al. [13] using third-grade polynomial regression. In their case, images were taken by a compact camera with white LED illumination and so appears also more practical than some of the earlier methods.

Spectral methods. Some works have incorporated non-visible light into computer methods for lycopene estimation, as already stated. The motivation for this is that better-discriminating, and generally richer, data for riper tomatoes may be accessible.

A linear correlation coefficient of R^2=0.96 between predicted and measured lycopene values was published by Polder et al. [14], using a hyperspectral camera with 256 spectral bands. A multispectral approach with 19 wavelengths using LED illumination by Liu et al. [15] gave a lower value of 0.94, but using a set-up more practical for non-laboratory conditions. Tihalun et al. [16] use both VIS/NIR spectrometer and chroma meter for Hunter L*a*b* representation of VIS. In con-

trast to other works, that paper used *transmitted* light passing through the tomato sample rather than reflected light. Results favoured the L*a*b* method: R^2=0.96 compared to R^2=0.85 with the spectrometer.

Discussion of the prior work. The non-destructive lycopene content detection methods considered above are presented in Table 1. The results suggest that non-destructive estimation of the lycopene content by optical sensors is viable. Five methods have R^2 higher than 0.95, of which, four are based on L*a*b* colour space. Multi/hyper spectral methods have an average correlation of R^2=0.916 compared to R^2=0.943 for the L*a*b* colour space methods.

Table 1: Comparison of previous methods with that proposed in this paper.

Lycopene detection result/parameter \ Authors	Best correlation R^2	Result notes	Linear correlation	Cross validation	Regression method	Colour space	Light interaction	Spectrum bands /space	Wavelength min	Wavelength max	Ground truth	No. of samples
Arias et al., (2000)	0.960	Chroma meter	N	N	Logarithmic	CIE Lab	Reflect.	a*	VIS	VIS	HPLC	38
Polder et al., (2004)	0.960	Line scan	Y	LOOCV	SIMPLS reg.	Hyperspectral	Reflect.	256	396	736	HPLC	37
Vazquez-Cruz et al., (2013)	0.985	Spectrophoto meter	Y	N	MLP (NN)	Hunter Lab	Reflect.	a*/b*	VIS	VIS	HPLC	48
Liu et al.,(2015)	0.938	Snapshot	Y	LOOCV	BPNN	Multispectral	Reflect.	19	405	970	SPM	162
Ye et al., (2018)	0.810	Snapshot	Y	N	Linear regr.	CIE Lab	Reflect.	$(a^*/b^*)^2$	VIS	VIS	SPM	60
Barrios et al., (2018)	0.985	Snapshot	N	N	Polyn.4gr	CIE Lab	Reflect.	a*	VIS	VIS	SPM	60
Tilahun et al., (2018)a	0.960	Chroma meter	Y	N	PLS regr.	Hunter Lab	Transmit.	a*/b*	VIS	VIS	SPM	180
Tilahun et al., (2018)b	0.850	Spectrometer	Y	N	PLS regr.	Hyperspectral	Transmit.	1160	500	1100	SPM	180
Mlynarik et al., (2022)a	0.977	Snapshot	Y	LOOCV	SNN	Multispectral	Reflect.	15	365	940	SPM	50
Mlynarik et al., (2022)b	0.959	Snapshot	N	N	Polyn.4gr	CIE Lab	Reflect.	a*	VIS	VIS	SPM	50

The success of the L*a*b* methods are probably due the a* parameter representing a green (chlorophyll) to red (lycopene) transition, reflecting a tomato's natural colour changes during maturation. Fig. 1, shows the relationship between a* and lycopene concentration using data captured for this paper (method described below). That is, an initial rapid transition from green to red as lycopene increases, followed by minimal change in a* thereafter. This demonstrates why a* alone can be successful, but also that it is not very discriminating for ripe tomatoes. In addition, hardware used for a* methods are well established off-the-shelf components with time-proven calibrations, compared to hyperspectral or multispectral systems which are usually bespoke with proprietary calibration methods.

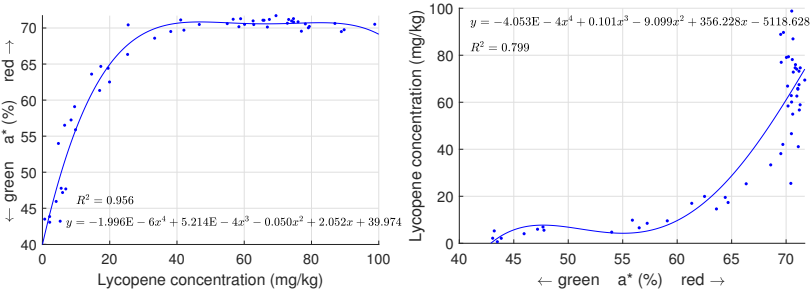

Figure 1: Measured relationship between a* and lycopene concentration. The graphs are identical with polynomial regression, but with axes reversed.

A higher R^2 value for a given regression might be an indicator of a superior fit to the data, but it can also be misleading in terms of achieving a useful model. For example, regression of measured a* vs ground truth lycopene concentration can be high as R^2= 0.96 or as low as R^2=0.80 depending on the somewhat arbitrary axis order (Fig. 1). Further, the regression offers no scientific basis to the underlying relationship. R^2 of a linear regression between estimated and ground truth lycopene is more robust due to its resilience against over-fitting (as is root mean squared error (RMSE)). Unfortunately, not all past methods provide such parameters for comparison.

In addition to accuracy, other important factors for real-world application are practicality, speed and cost. The highest R^2 in L*a*b* methods are detected using multiple points around the sample relying on close proximity of the sensor (e.g. [5, 11]). Such a sampling technique is less practical than a single distant snapshot for high-throughput, high-speed sorting applications. Hyperspectral and multispectral techniques with more bands might increase the complexity of the system further. Therefore, the requirement of our method (and some others) for specialised illumination must be balanced against its benefits of more robust data capture.

3 Multispectral method for lycopene estimation

For this research, multispectral light reflections in 15 bands between 365nm and 940nm were used to investigate the precision of the method and its practicality for use in a controlled but non-contact industrial environment. The aim was to attain robustness and high correlation of predicted and measured lycopene content, especially for fully ripe tomatoes, while using commercially available devices that can easily be deployed in industry. The wavelength range was selected based on the assumption that a multispectral system consisting of more than three bands, covering both the full VIS spectrum and beyond, should contain more information than a system just utilising RGB sensor information converted to L*a*b*. That is, the L*a*b* data comprise a subset of the broader multispectral data and so should not exceed it in performance.

In this paper, multispectral data capture is optimised in the following ways. (1) Tomatoes were illuminated by dome lighting to avoid shadows and specular reflections. (2) The size and hardware construction were chosen to ensure uniform intensity over the entire fruit 3D surface. (3) The tomato was imaged from four sides to avoid situations where the red pigment is not evenly established during growth. While this arrangement might have limited direct applicability, the aim is to establish a robust baseline on which to build upon in future research.

Experiment: methods and materials. Fifty cultivar Saluoso tomatoes were harvested in late-autumn from a hydroponic greenhouse in south-east Slovakia. They were selected randomly, but covered a complete range from fully green to fully red. A multispectral image was captured (see below) for each tomato sample. Each sample was then blended within an hour and dissolved in hexan-etlylen-aceton followed by spectrophotometric absorbance measurement at 503nm, in accordance with the method of Anthon and Barrett [17]. This process allowed the acquisition of a ground truth baseline from which comparisons could be made. One sample was later removed due uncertainty during dissolution.

Multispectral images were captured by a Basler Ace monochromatic and near infrared area-scan camera. For each case, a series of LEDs in the range 365nm to 940nm were used to illuminate the sample in a

bespoke Technomedia dome with 340mm inner diameter. The system was calibrated with a spectralon target plate at seven points to ensure uniformity of image intensity between each wavelength.

Images were then segmented using basic thresholding functions in Halcon software. Next, image processing was split into two paths. (1) Convert the three images corresponding to RGB bands (478nm, 520nm, 635nm) to the L*a*b* colour space to calculate an average pixel intensity of a* for correlation with lycopene concentration. (2) Average segmented image intensities were fed into a shallow neural network (SNN), with five hidden layers, trained using the MATLAB `fitnet` function to map the multispectral data to measured lycopene values.

Tomato surface area involved in computation. Lycopene is not distributed evenly inside tomatoes, but is almost four times more concentrated in the skin compared to the pulp, and five times higher than the seeds [18]. Further, different parts of a tomato's surface may be more mature than other parts. The multispectral images were therefore taken from four sides: stem, bloom, left and right. Results are shown in Table 2. These R^2 regression results confirm the hypothesis that larger coverage improves correlation.

Polynomial regression of	Side of tomato for taking image				
	All	Bloom	Stem	Left	Right
4 grade	0.9557	0.9531	0.9465	0.9508	0.9336
3 grade	0.9467	0.9429	0.9360	0.9431	0.9438
2 grade	0.8847	0.8779	0.8735	0.8830	0.8682
Average above	0.9290	0.9246	0.9187	0.9256	0.9152
Log. regression	0.8965	0.8926	0.8862	0.8943	0.8760

Table 2: R^2 correlation between a* and ground truth lycopene content for various sides of sample. Logarithmic, 2nd, 3rd and 4th grade polynomial regressions are shown.

Selected spectra and wavelength bands contribution. The green colour of unripe tomatoes is due to the prevalence of chlorophyll. During ripening, the synthesis of lycopene results in a red colour. Lycopene has a carotenoid molecular structure of eleven double bonds, allowing it to absorb energy from UV light between 270 and 310nm and blue and green light between 350 and 530nm [19]. In the proposed method therefore, this range is covered with seven spectral bands from 365 to 520nm. This is in addition to three wavelength bands in red spectra to capture the green to red colour shift. In total 15 wavebands were included, including NIR.

In Fig. 2, the measured average intensity of each waveband is plotted as a function of ground-truth lycopene concentration. Polynomial regression lines are also shown for ease of comparison. The figure shows several wavelengths with similar shape, suggesting little benefit of including them all. However, about seven different trends can be recognised. For a well-designed neural network, during training, the weights will become optimised to exploit these trends.

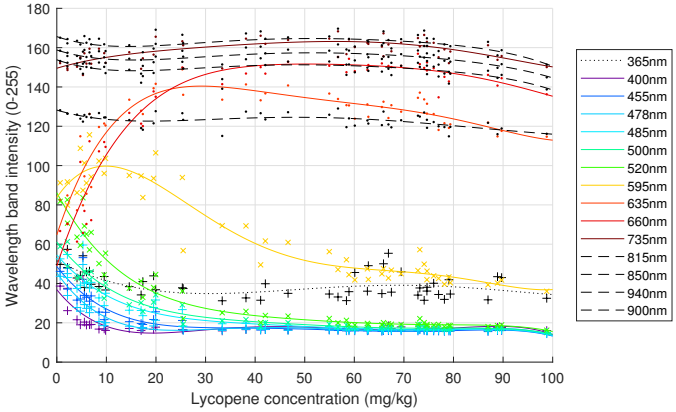

Figure 2: Averaged pixel intensity of reflected light as a function of lycopene for all wavelengths. [Colour coding approximately matches wavelength. "+": ultraviolet/blue, "x": yellow/green, "·": red/infrared.]

Shallow Neural Network (SNN). The additional information available from multispectral data was incorporated using Levenberg-Marquardt backpropagation SNN with 5 hidden layers. This approach is known to better model the non-linear interaction of sparse data. Modern methods for computer vision typically use convolutional neural networks (CNNs). However, that is deemed unnecessary here since the inputs are single values corresponding to mean intensity measurements for each wavelength (i.e. there is little benefit from setting entire images as inputs, as expected by most CNN architectures). In future work, it might be possible to use CNNs in order to incorporate potentially useful spatial information.

To investigate the influence of the various wavebands on the appearance of lycopene, an SNN was trained for all possible band combinations using identical settings. In addition, one more input to the SNN was added: the physical size of the tomato sample as a 16th possible input. The motivation for this is that, as lycopene is more highly concentrated near the surface, the physical size may affect average concentration levels of the sample. As presented below, the best prediction was, indeed, achieved with that additional input.

For evaluation, the leave one out cross validation (LOOCV) method was used. Given a sample size of 49 therefore, 49 training sessions were performed for each of 65,535 possible combinations of wavebands from 1 to 16 bands. Fig. 3 shows the general effect of the number of bands considered (1,2,...16) in terms of performance.

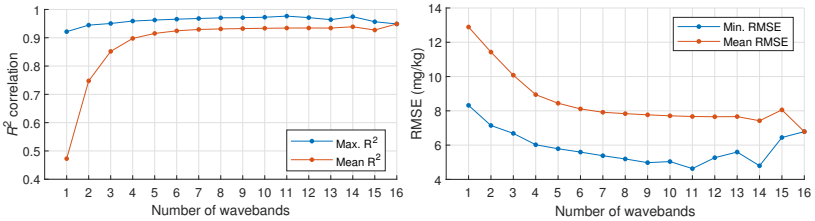

Figure 3: SNN performance expressed in maximum and average R^2 correlation (left) and minimum and average RMSE (right) of prediction against number of input wavelength bands.

Multispectral LOOCV linear regression correlation reached a maximum of R^2=0.9765 for lycopene prediction and measured ground truth concentration. This corresponds to RMS error of prediction of 4.63 mg/kg. A combination of 11 bands gave this result (all those in the legend for Fig. 2 except 485nm, 520nm, 635nm, 850nm).

It was found that the SNN performance does not improve when the number of input bands is above about eight. This might be due to the introduction of noise with additional bands with very similar shape or due to the model over-fitting. Therefore, although the 11 wavebands in the optimal SNN mentioned above had best correlation in experiments, it is likely that almost equally good outputs are possible with fewer (not necessarily identical) inputs.

Discussion. Our method allowed us to explore both the multispectral and the L*a*b* approaches. At best, we found that fitting a* against lycopene concentration using 4th grade polynomial regression gave R^2=0.9557. While this sounds promising, the a* value rapidly converges with moderate lycopene concentration, meaning the regression curve has limited use above certain maturity levels. This problem is also apparent in some other research that focuses on L*a*b* space. Further, high-grade polynomials such as this are widely known to over-fit and should be interpreted with care.

As an alternative to the above approach, where polynomial fitting might be somewhat arbitrary, we have also trained SNNs with varying numbers of hidden layers for all possible combinations of the wavelengths and sample size. Through trial-and-error, it was found that results improved with the number of hidden layers up to about 5, beyond which, little improvement was obtained. For this reason, only results from SNNs with exactly five hidden layers are presented. Results show that the stability and prediction of correlation increase with the number of wavebands, as hypothesised. Additional bands, including those outside the visual spectrum, have proven their contribution to model robustness and preciseness.

The results from both previous works and our own, are shown in Table 1. This indicates that the performance of our method is comparable to others, while maintaining a more reproducible approach and application of cross-validation, which not all others do.

4 Conclusion

While previous research has shown promise for lycopene concentration estimation using computer vision, this research offers a more robust grounding with detailed experiments in controlled conditions. This demonstrates what may be possible using intensity analysis at a range of wavelengths in a laboratory setting, which can be reproduced with minimal cost. The limitations of L*a*b* space are demonstrated and it is shown how our multispectral approach goes some way to overcome these using neural networks. Future work will aim to investigate how the approach can be extended to operate in an agricultural setting.

References

1. STATISTA.com: Vegetables: Worldwide., https://www.statista.com/outlook/cmo/food/vegetables/worldwide, [Accessed: 11 January 2023].
2. "Lycopene: modes of action to promote prostate health," *Archives of Biochemistry and Biophysics*, vol. 430, no. 1, pp. 127–134, 2004.
3. S. Przybylska, "Lycopene – a bioactive carotenoid offering multiple health benefits: a review," *International Journal of Food Science and Technology*, vol. 55, pp. 11–32, 2019.
4. World Food and Agriculture – Statistical Yearbook 2021., https://doi.org/10.4060/cb4477en, [Accessed: 11 January 2023].
5. R. Arias, T.-C. Lee, L. Logendra, and H. Janes, "Correlation of lycopene measured by hplc with the L,a,b color readings of a hydroponic tomato and the relationship of maturity with color and lycopene content," *Journal of Agricultural and Food Chemistry*, vol. 48, no. 5, pp. 1697–1702, 2000.
6. J. Wieme, K. Mollazade, I. Malounas, M. Zude-Sasse, M. Zhao, A. Gowen, D. Argyropoulos, S. Fountas, and J. Van Beek, "Application of hyperspectral imaging systems and artificial intelligence for quality assessment of fruit, vegetables and mushrooms: A review," *Biosystems Engineering*, vol. 222, pp. 156–176, 2022.
7. W. Wardencki, P. Biernacka, T. Chmiel, and T. Dymerski, "Instrumental techniques used for assessment of food quality," *Proceedings of ECOpole*, vol. 3, 2009.
8. W. W. Fish, P. Perkins-Veazie, and J. K. Collins, "A quantitative assay for lycopene that utilizes reduced volumes of organic solvents," *Journal of Food Composition and Analysis*, vol. 15, no. 3, pp. 309–317, 2002.
9. A. Hussain, H. Pu, and D.-W. Sun, "Measurements of lycopene contents in fruit: A review of recent developments in conventional and novel techniques," *Critical reviews in food science and nutrition*, vol. 59, no. 5, pp. 758–769, 2019.
10. M.-J. Villaseñor-Aguilar, J.-A. Padilla-Medina, J.-E. Botello-Álvarez, M.-G. Bravo-Sánchez, J. Prado-Olivares, A. Espinosa-Calderon, and A.-I. Barranco-Gutiérrez, "Current status of optical systems for measuring lycopene content in fruits," *Applied Sciences*, vol. 11, no. 19, p. 9332, 2021.
11. M. Vazquez-Cruz, S. Jimenez-Garcia, R. Luna-Rubio, L. Contreras-Medina, E. Vazquez-Barrios, E. Mercado-Silva, I. Torres-Pacheco, and R. Guevara-Gonzalez, "Application of neural networks to estimate carotenoid content during ripening in tomato fruits (solanum lycopersicum)," *Scientia Horticulturae*, vol. 162, pp. 165–171, 2013.

12. X. Ye, T. Izawa, and S. Zhang, "Rapid determination of lycopene content and fruit grading in tomatoes using a smart device camera," *Cogent Engineering*, vol. 5, no. 1, p. 1504499, 2018.

13. A. Gastélum-Barrios, J. García-Trejoo, G. Soto-Zarazúa, G. Macías-Bobadilla, and M. Toledano-Ayala, "Portable system to estimate ripeness and lycopene content in fresh tomatoes based on image processing," in *Int. Eng. Congress*, 2018, pp. 1–5.

14. G. Polder, G. Van Der Heijden, H. Van der Voet, and I. Young, "Measuring surface distribution of carotenes and chlorophyll in ripening tomatoes using imaging spectrometry," *Postharvest biology and technology*, vol. 34, no. 2, pp. 117–129, 2004.

15. C. Liu, W. Liu, W. Chen, J. Yang, and L. Zheng, "Feasibility in multispectral imaging for predicting the content of bioactive compounds in intact tomato fruit," *Food Chemistry*, vol. 173, pp. 482–488, 2015.

16. S. Tilahun, M. H. Seo, I. G. Hwang, S. H. Kim, H. R. Choi, C. S. Jeong *et al.*, "Prediction of lycopene and β-carotene in tomatoes by portable chromameter and VIS/NIR spectra," *Postharvest biology and technology*, vol. 136, pp. 50–56, 2018.

17. G. Anthon and D. M. Barrett, "Standardization of a rapid spectrophotometric method for lycopene analysis," in *X International Symposium on the Processing Tomato 758*, 2006, pp. 111–128.

18. R. K. Toor and G. P. Savage, "Antioxidant activity in different fractions of tomatoes," *Food Research International*, vol. 38, no. 5, pp. 487–494, 2005.

19. H. Hashimoto, C. Uragami, and R. J. Cogdell, "Carotenoids and photosynthesis," *Carotenoids in nature*, pp. 111–139, 2016.